Email Archiving Stewardship Tools Workshop

Harvard Library Report
July 2016

Prepared by Chuck Patch

HARVARD LIBRARY

Prepared by Chuck Patch, Cultural Information Management Consulting

Reviewed by Wendy Marcus Gogel, Harvard Library and Grainne Reilly, Library Technology Services, Harvard University

Citation:
Patch, Chuck. 2016. Email Archiving Stewardship Tools Workshop. Harvard Library Report. http://nrs.harvard.edu/urn-3:HUL.InstRepos:28682573.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

On March 2, 2016, Harvard Library brought together a group of leading practitioners in the field of email archiving for a two-day workshop to assess the ways the community could work together on the growing selection of tools developed in the past few years for managing the exploding volume of this archetypal form of communication. Inspired by the 2015 *Archiving Email Symposium* hosted by the Library of Congress and the National Archives and Records Administration, and including a number of the same participants, the workshop was intended to forward the work set forth at that event. As proposed by workshop leader Wendy Gogel, Manager of Digital Content and Projects for Harvard Library Preservation Services, the workshop set out to achieve four goals:

- To foster the expanding email archiving community
- Share updates on current work
- Expose the Harvard Library community to the issues involved in email archiving
- Identify needs for upcoming work and future directions

The participants included many of the principal scholars, organizers, and developers in the field of email archiving. As Gogel pointed out during the opening session, in simply bringing the group together, which in addition to eight members of Harvard's Library staff included representatives from Stanford University, the University of Illinois Urbana Champaign, the Smithsonian Institution, the Library of Congress, the University of North Carolina Chapel Hill, MIT, as well as the Director of Technical Services for Archivematica, an independent non-profit software developer, they had achieved the first of the workshop's goals.

Over the next two days, the workshop moved through an agenda structured to achieve the remaining goals. Formal presentations that included analyses of typical email archiving strategies, and overviews of the existing management tools (all of which had been developed by participants of the workshop) established the current state of the field. From there, the participants looked at current efforts to objectively compare tools so that practicing archivists could make intelligent software choices.

During the workshop, the Harvard Library community was able to engage with the participants several times, including the first afternoon of formal presentations in the conference room, an evening mixer and dinner, and an open presentation at the Lamont Forum room on the second day.

The heart of the workshop consisted of spirited discussions in response to use cases intended to highlight gaps in tool functions and processing workflows, and to identify potential ways that the community could work together to fill them. A critical turning point in the session occurred when Christopher Prom suggested that each of the tools represented in the use cases have different strengths and are good at different tasks. For example:

- ePadd is very good at appraisal
- DArcMail is very good at preservation and the wrapping up of a digital object using an xml standard

- The EAS tool is good at metadata creation, and the pushing out of data
- Archivematica does a good job providing preservation services for attachments and the messages themselves

The question arose as to whether the community would agree to a single workflow or whether the need is to agree to a common exchange format to enable the use of multiple tools to fulfill workflows. The participants came to a vital consensus: there is no such thing as a standard workflow for processing email.

Where the workshop organizers and many in the community had begun with the hope of establishing a singular, linear (and potentially monolithic) workflow that would commonly be used by all, it became apparent that archivists require a more flexible, "mix-and-match" approach to tool selection and the design of workflows. Rather than assuming that processing can be done end-to-end in a certain order, with a single tool, tools may be used based on their strengths in various combinations depending on the individual institution's needs. Archivists may wish to use the same tool for more than one function, exporting data in mid-workflow for processing in another tool that performs a certain function more suited to their needs, and then re-importing their data back into the first tool for continued processing.

The potential importance of this approach was immediately recognized by the participants, who then went on to identify three areas in need of further effort. For each need, they defined immediate goals to be pursued voluntarily by individual workshop contributors, including:

1. *Explore a common exchange format to enable the use of multiple tools to fulfill workflows*

   During the workshop, the group began to think about what each tool would need to know and do to interoperate with data from the other tools. To solve this problem in a practical way, the group started a spreadsheet for gathering information about which data is required by each of the systems. This represented a first attempt to explore the exchange of metadata and content between email archiving tools based on their inputs and outputs.

   Following the workshop, Harvard hired Artefactual to refine this spreadsheet into a workable framework for information about tool requirements regarding exchange of email metadata and content; and then to collect the data for the spreadsheet.

2. *Explore methods of sharing lexicons that can be used by the multiple tools*

   The need for sharing lexicons across multiple tools was identified as a critical issue during the workshop. Methods to accomplish this are being addressed by an ePADD Lexicon Working Group, led by Josh Schneider at Stanford, and including Kari Smith (MIT) and Glynn Edwards (Stanford). The Working Group held their first meeting in June 2016. Several participants shared the lexicons that are being developed and tested at their institutions. The ePADD project plans to post them on the project website.

3. *Develop tools for identification and validation of sustainable formats for email*

Kate Murray volunteered to take this idea back to the Library of Congress as a distinct area to pursue as part of LC's work on format standards.

In the end, the workshop solidified the community's interest in working together to solve the problems they face. The group recognized that a sustainable approach to email archiving requires tools with the flexibility to be combined for diverse workflows and was able to agree on, and prioritize, the initial work (numbers 1 to 3, outlined above). Each of these areas sets a promising direction for future collaborative work.

Information about the workshop and its results were disseminated subsequently:

- On March 9, 2016, Harvard Library released an article in the online newsletter: Email Archiving Stewardship Workshop at http://library.harvard.edu/03092016-1642/email-archiving-stewardship-workshop. (*See* pp. 71)
- On March 21, 2016, Kate Murray and Wendy Gogel contributed a summary of the Harvard EAST Workshop to a discussion about email archiving as part of the National Digital Stewardship Alliance (NDSA) Standards and Practices Working Group phone call (http://ndsa.org/working-groups/standards-and-practices) that included Mellon Foundation representatives.
- On May 10, 2016, Kate Murray posted on The Signal — the Library of Congress digital preservation blog: O Email! My Email! Our Fearful Trip is Just Beginning: Further Collaboration with Archiving Email at http://blogs.loc.gov/digitalpreservation/2016/05/o-email-my-email-our-fearful-trip-is-just-beginning-further-collaborations-with-archiving-email. (*See* pp. 73)

WORKSHOP PARTICIPANTS

- **Glynn Edwards**, Head, Technical Services Division, Dept. of Special Collections & University Archives, Stanford University
- **Riccardo Ferrante**, Information Technology Archivist & Digital Services Program Director, Smithsonian Institution Archives
- **Franziska Frey**, The Malloy-Rabinowitz Preservation Librarian; Head of Preservation and Digital Imaging Services, Harvard Library
- **Andrea Goethals**, Manager of Digital Preservation and Repository Services, Harvard Library
- **Wendy Gogel**, Manager of Digital Content and Projects, Harvard Library
- **Skip Kendall**, Sr. Collection Development and Electronic Records Archivist, Harvard Library
- **Cal Lee**, Associate Professor, School of Information and Library Science, University of North Carolina, Chapel Hill
- **Anthony Moulen**, Library Technology Architect, Harvard University Information Technology
- **Kate Murray**, IT Specialist (Audio-Visual Specialist), Technology Policy Directorate, Library Services, Library of Congress
- **Chuck Patch**, Facilitator/Consultant
- **Tricia Patterson**, Digital Preservation Analyst, Harvard Library
- **Christopher Prom**, Associate Professor of Library Administration, University of Illinois, Urbana-Champaign
- **Grainne Reilly**, Sr. Digital Library Software Engineer, Harvard University Information Technology
- **Justin Simpson**, Director, Archivematica Technical Services, Artefactual
- **Kari Smith**, Digital Archivist, Institute Archives and Special Collections, MIT Libraries
- **Randy Stern**, Director, Systems Development, Harvard University Information Technology

WORKSHOP PROCEEDINGS

Wednesday, March 2, 2016

WELCOME AND INTRODUCTION BY FRANZISKA FREY AND WENDY GOGEL

W. Gogel welcomes panel and introduces F. Frey. F. Frey discusses the past history of dealing with email on an institutional basis, and discusses the potential for moving from an inward institutional focus to a broader community.

W. Gogel reviews the history of the group, many of whom have collaborated in the past, and a number of whom were on an SAA panel in 2015 on "Email Archiving in a Curation Lifecycle Context." The impetus and inspiration for this workshop was the Archiving Email Symposium (AES) and workshop co-hosted by the Library of Congress and the National Archives and Records Administration in June 2015. On June 2, the AES shared information about the state of practice in accessioning and preserving email messages and related attachments for an audience of approximately 150 people. (For more information, see: http://www.digitalpreservation.gov/meetings/archivingemailsymposium.html?loclr=blogsig). On June 3, there was an informal workshop with a subset of participants to discuss issues and challenges identified during the Symposium in order to better define the gaps in our tools, processes and polices for archiving email collections. - http://blogs.loc.gov/digitalpreservation/2015/07/we-welcome-our-email-overlords-highlights-from-the-archiving-email-symposium/

Harvard Library would like to continue to build on this work.

Workshop Goals

W. Gogel reviews the goals for the workshop:

- to foster the expanding email archiving community
- share updates on current work
- expose the Harvard Library community to the issues involved in email archiving
- identify needs for upcoming work and future directions

She points out that 3 of the 4 goals for the meeting have been achieved merely by getting the group together to exchange ideas.

INTRODUCTION BY CHUCK PATCH

C. Patch begins by asking each member of the group to introduce themselves and describe something that they hope to come away with in the meeting. Among the desires expressed:

- How to characterize email. Is it a thing or a collection of things?
- The sustainability of the tools, and how they'll interact. It is noted that building out the existing suite of Open Source tools, and adding to them will almost certainly aid with sustainability.
- Hope that the group will compare their approaches to archiving email, and try to establish whether there is a workflow that can be applied in most circumstances, or if the real issue is interoperability of the tools, and how that would work.
- The importance of tools that can handle very large scale collections.
- The hope that the workshop will help provide direction to the field, and to Harvard in particular as it considers its own direction.
- The hope that the results of the group's efforts will help lay the groundwork for a wider scope of PIM

Basic housekeeping issues were explained, and the agenda was reviewed. Other Harvard participants form the Library would join the workshop for some of the demonstrations by participants, a networking mixer, and a dinner on the first night. The workshop structure was intended to identify gap areas, and the common community needs for addressing those gaps. The workshop would look at the current state of the field from different perspectives, as presented by members of the group in presentations, including:

- An examination of processing workflow considerations
- Updates and demos on major tools in the field
- A discussion attempting to identify gap areas, or desired functionality, to be kicked off by a Use Case in the final hour of the day

It was explained that during the second day, a longer discussion was to take place, based on the compilation of a list of topics that each participant was invited to add to between the final product demo and the final hour of the day.

Participants were asked to note ideas for discussion topics during the demonstrations, which represented a slightly more formal phase of the workshop and included attendance by additional Harvard members.

The Harvard EAST Wiki was presented and participants were also urged to contribute to the linked forum discussions, and to email documents and slides to Tricia Patterson for posting:

https://wiki.harvard.edu/confluence/display/digitalpreservation/Harvard+EAST+Workshop

PRESENTATIONS AND DEMOS

THE BULK OFTHE AFTERNOON WAS DEDICATED TO PARTICIPANT'S PRESENTATIONS AND DEMONSTRATIONS OF CURRENT EFFORTS AND TOOLS USED FOR ARCHIVING EMAIL

1. Christopher J. Prom – University of Illinois Urbana-Champaign (UIUC) Computer Science meeting and an "ideal" email processing workflow

2. Glynn Edwards - ePADD demo
3. Riccardo A. Ferrante - Latest Tweaks to DArcMail
4. Skip Kendall - EAS demo
5. Cal Lee - Email processing in BitCurator context/ possible implications of BitCurator access tools
6. Justin Simpson - SFU workflow for ingesting email from Zimbra server to Archivematica
7. Kate Murray - Update on the collaborative Lifecycle Tools for Archival Email Stewardship (a.k.a "The Chart").

CHRIS PROM — FITTING EMAIL INTO AN APPRAISAL, ACCESSIONING, PROCESSING, DISCOVERY, AND DELIVERY WORKFLOW

Chris began by noting that the problem of sorting out the process of managing email preservation has continued to grow as a concern in the archival community. His presentation focused on the ways his repository handled the processing of email, and how these processes might be generalized to other repositories.

He began by comparing the processing of digital and analog materials, noting the points at which they are the nearly the same, similar, or different. Where the systems diverged radically, as in the initial processing steps, where digital materials would require a computer setup using a variety of software, there was an impact on the succeeding step. For example, although the desire is to use the same descriptive system for both analog and digital material, the processing steps altered the descriptive workflow process, posing a challenge.

Three steps in the workflow process were different from one another:

- Processing (Processing Tables <-> Processing Computer(s)
- Collections Storage (Collections Stacks <-> Preservation Repository)
- Delivery Systems (Reading Room/Remote Services <-> Tools we build and Tools users bring)

The last point highlighted the importance of tools that visitors and users brought, because there were always going to be tools that the repository is unaware of that might better server user's needs.

The point was made that both analog and digital materials used the same discovery system. Chris pointed out that this was the case at the UIUC, and for other archives as well, using tools such as ArchiveSpace. He did not believe that this would necessarily always be the case. At some point it is possible that they would have separate systems that would run in parallel, but if that were the case, they would need better tools for metadata interoperability or search tools that ran across multiple systems.

The management of analog and digital materials at UIUC and other repositories is a balancing act between standards, tools, plans, available personnel, and the infrastructure in place. Designing tools that fit into the decision-making process at a given institution is challenging, because these processes are likely to be different among institutions.

An important issue in his repository, and probably many others, is that the workload is very heavy, and that they attempt as much as possible to do things only once and be done with it. A critical part of workflow design is to find the right balance between manual and automated processes.

These are baseline decisions that affect handling of all digital materials. From the perspective of the UIUC archive, email constitutes just one type of born digital material.

A result of this approach is that regardless of the type of material, they have developed a single SIP, or AIP packet, to represent the entirety of the material from the creator of the material. Thus there will be one Collections Level entry in their descriptive system, and one digital object record describing the digital materials in that collection. Similar decision on processing have been reached by archives at Michigan, Kent State and Yale.

Accruals over time are handled by creating another AIP, and creating a link between the AIPS pertaining to the same collection.

Basic assumptions

- Processing regimen is very light-weight.
- Loosely coupled relationship between the records in various systems, using the record ID as the linkage.

MICHAEL HART COLLECTION

As an example of their workflow process, Chris walked through the processing of the Michael Hart Collection. (PowerPoint presentation at http://bit.ly/21L20ZT .)This collection includes many things, a component of which is email.

- They took forensic images of several of Hart's computers, and other media found in his house when they brought the collection into the repository.
- The unprocessed material was put into a "holding tank," which is a 20TB share in the University's share network share system. To the archivist, it appears as the "unprocessed" subfolder in the "UA" share. It stays there until they "decide what to do with it." *Slide 8 (*Figure **1***).*

Figure 1 Prom presentation, Slide 8

- An accession record is created that includes all of the material in the incoming collection, and this record is then made public in their holdings database. The born digital materials are alluded to in the comment. *Slide 9 – 10* **(Figure 3)**

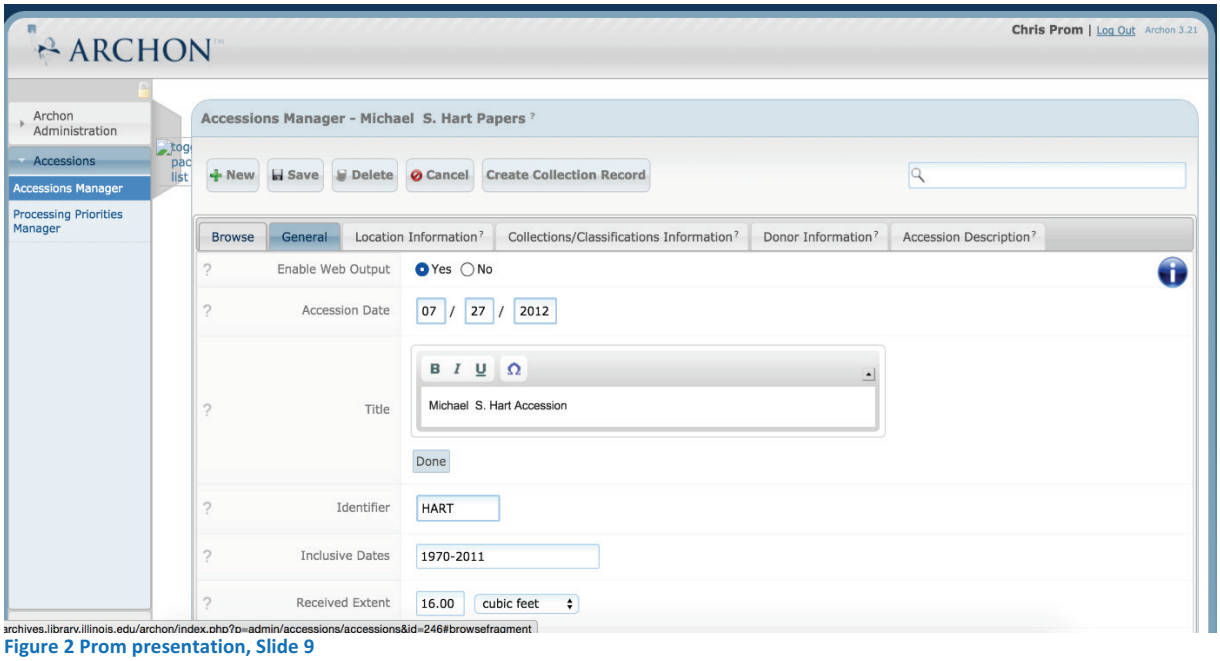


Figure 2 Prom presentation, Slide 9

- During this period, they are thinking of this entire set of material they will be processing in terms of the AIP, in particular keeping the content information (the digital files they acquired), the descriptive, and preservation information in the appropriate systems.

Figure 3 Prom presentation, Slide 10

- All of the born-digital material would be represented by one catalog system, in one record kept in their collections management system, which would then be placed in an AIP as represented in *slide 11*. **(Figure 4)**

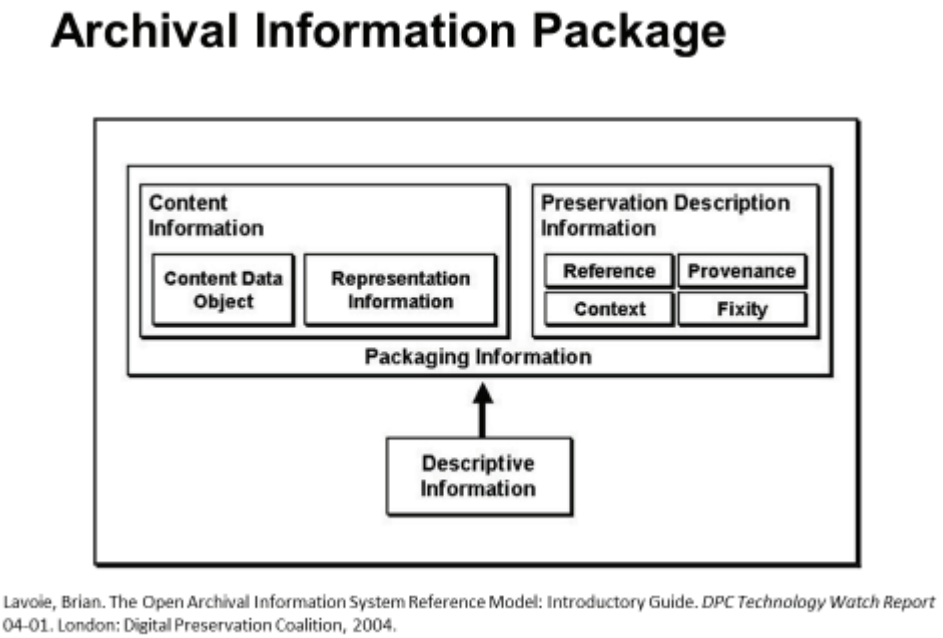


Figure 4 Prom presentation, Slide 11

The final result is represented in *Slide 12* **(Figure 5),** which indicates that the original digital material is placed in a preservation folder, while access copies are kept in online files, for material that has no

privacy or restriction concerns, and in a near-line folder that can only be accessed by the archivist. Preservation description information is stored in a parallel folder tree. All of these are kept in a folder tree that is labeled with the collection number.

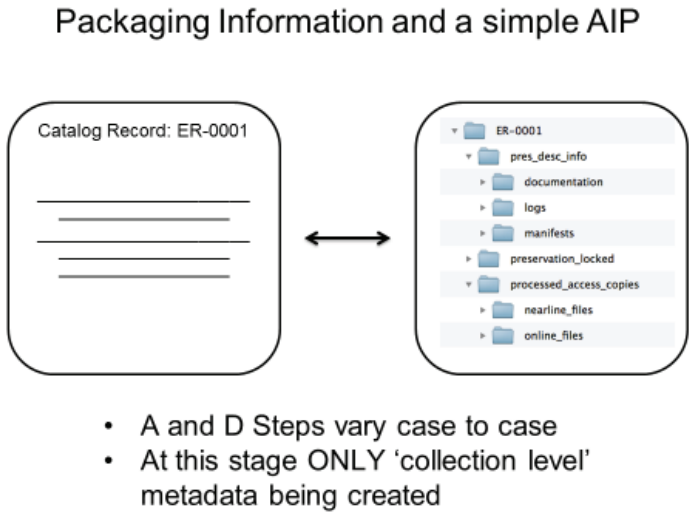


Figure 5 Prom presentation, Slide 12

This process took several months for the Hart Collection. During this time, they converted the email store to MBOX format. When complete, the description in their collections management systems' online interface describes the entirety of the collection, with a link to born digital materials

The files that they have processed get staged for upload to a local preservation repository. Although there is much variation in the implementation of repositories among different archives, the basic idea here is that they put everything into a folder, stage it for upload in a structure that uses a collection ID for linking back to the collection record. The local repository was created to manage bulk archival and library collections. The system used is bespoke, but Chris feels that they would have been better served by acquiring Archivematica.

One problem with their current arrangement is that all of the descriptive data in their Collections Management System is duplicated into the local preservation repository. Nonetheless, they can view and download the files from the system ("Medusa") any time they need via a web interface. They are already able to give access to some of the email files, but it's at the MBOX level only, so the user needs to download this and "do what they will" with it.

Their next steps – *Slide 22* **(Figure 6)** – include investigating ways in which they can take greatest advantage of derived, automatically generated metadata as possible. They are starting with photographs because they present a more frequently used set of metadata schemas, implemented in

the file headers of photographs. Another step will be to examine the extent to which different media types will require different retrieval and discovery interfaces.

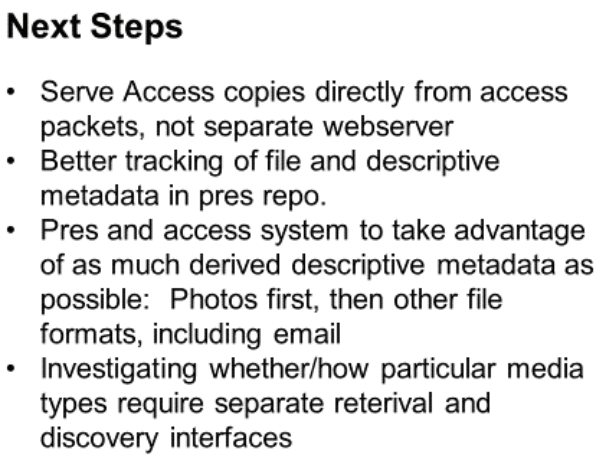


Figure 6 Prom presentation, Slide 22

Takeaways include the fact that their efforts have been concentrated thus far in getting high level control of both born digital and analog material. This is a necessary precursor to establishing better treatment for specific types of media. They also need ways to automatically extract series, file and item metadata to store and use it in a discovery system.

Finally, their system works best when they are able to swap in and out tools as necessary – an aspect of their situation he mentioned near the beginning of the talk, and also seemed to allude to when mentioning the need to have users bring tools they know of – and that the archivists do not -- into play when working with the collections.

QUESTIONS FOR CHRIS

Wendy Gogel brought up the issue of synchronization problems Chris alluded to when discussing the duplication of metadata during ingest. Chris pointed out that this synchronization problem is repeated with many of the tools that they use, and has driven many repositories to adopt a practice that they use, which is to treat the digital objects in a wholly separate system and not worry about the metadata for them in the archival management tool. Chris would prefer that users rely on the tools for digital information to develop the metadata, and have that propagate back to the archival discovery tool, but that is not possible yet.

Kari Smith described MIT's handling of the synchronization problems. In their case, they create the AIP at the point of accession. Using Archivematica, they create multiple SIPS that are then ingested into the repository. Their process involves multiple round-trips for the object metadata records that refines the

records over time, and reduces the synchronization problems. The ability to do this is a result of planned and new functionalities in Archivematica (version 1.6) and ArchivesSpace (version 1.5)

Someone asked about the origin of the "holding tank." According to Chris, it was developed to handle the problem of widely dispersed materials of any given collection. By gathering all of the materials from various computers and storage media into a single folder, linked by ID, they were able to make processing more orderly. The shared storage is managed by the Archives group. MIT has a similar facility by using four networked storage areas – one is for Transfers. Kari described how they began by examining the materials in unprocessed collections in this holding Transfer area, and determining how to process different segments depending on issues such as rights.

Kari asks if they provide instructions to end users downloading the MBOX files for how they can access the materials in the file (e.g. 'you can use Thunderbird to access individual emails.') They have not done this at UIUC partially because they're concentrating on getting collections level access. Although UIUC currently keeps access copies and preservation copies in separate stores, the plan is eventually to serve the online version off the top of the preservation repository.

GLYNN EDWARDS — EPADD

The PowerPoint provided in the group Wiki : "archival stewardship_HUL Th.pptx," located on the workshop wiki at: http://bit.ly/1UpXpO9 includes notes for the slides. The following notes summarize her talk, with greater concentration on material that is not included in the PowerPoint presentation. This includes a live demo of the newest version of ePadd, which is scheduled for release in June, 2016. The current version, 1.0, was released in the summer of 2015, with an interim version that included a few new features (which she discussed) released in the Fall of 2015.

The project is a joint collaboration of five institutions, including Harvard, UC Irvine, Metropolitan New York Library Council, and the University of Illinois

The project benefitted from a number of funding sources, which allowed them to begin work on certain features of the system before they knew that the current, 3 year IMLS grant had been approved, including the Pilot Discovery Platform, and it also allowed them to hire a UI firm. A programmer with the UI firm was able to develop a fine-grain entity extractor for the system, before the current grant was initiated, which will be part of the 2016 release. Most of her talk, however, was devoted to the first 2 releases of the software.

Their effort has been to allow a creator or a donor, as well as archivists and repositories to search for specific features in an email archive that they may want to restrict, or not transfer. Use cases have driven the development of some of the features. The most important of these is the ability to search for restricted data. For example, they can do regular expression searching, and the back-end is editable for whatever pattern is desired.

After the archives had acquired the archive of Richard Fikes, and before they had made it available to the public, he submitted a list of 300 correspondents whose emails he did not wish to be made public. To locate all these correspondents' email addresses, they developed a utility for importing csv files.

The Community Manager for ePadd has developed an overview video that explains how the system works for the uninitiated that is on the front page of the ePadd website (https://library.stanford.edu/projects/epadd .)

She presented a list of issues that are currently being addressed for the June release. These issues are presented on their Github page: https://github.com/ePADD/epadd/issues. Note that this aspect of her presentation is not included in the PowerPoint presentation on the workshop wiki sit. Among these issues are:

- Lots of bug fixes
- Advanced search
- A new version of the Named Entity Recognizer (NER)
- A new UI, developed by Lollypop, the developer of the current interface.

All the updates will be available through the Github site (https://github.com/ePADD/epadd), which will include release dates for the new features. In addition, a much fuller list of issues, and feature developments are available via the Github site.

These new issues were (and are being) developed in response to needs that were identified with the previous release. They were voted on by the collaborators and ranked in priority order. As Glynn notes in her PowerPoint presentation:

> *During processing we wanted to provide the ability to:*
>
> *- do pattern searching so we could review PII information*
>
> *- Flag items for transfer (or not)*
>
> *- Flag and annotate restricted materials*
>
> *We also included configuration files (which are editable and persistent across repository) such as:*
>
> - *Named entity kill-lists*
> - *multiple lexicons, including: regular expression lists*

The project is very fluid: part research, part testing, and part development. They are exploring:

- *Preservation.* Not part of their original NHPRC grant (or release), but very important to the wider community. This will require many conversations with their partners and a much closer collaboration with their digital library systems and services department (DLSS). Up until this point DLSS has not been involved with the project.

- *Cross-collection searching*
- *Cross-institutional searching for discovery metadata*
- *Linked Open Data* (either extracting or publishing)
- *Social Networking*
- *Visualizations*
- *Annotation Management*.

They have no plans to incorporate records management, but will seek input from them in year one of the release by hosting a conference with records managers.

The live demo included a look at some of newer versions of screens depicted in *Slide 6 and Slide 7* (**Figure 7, Figure 8**) using a subset of about 4000 messages from the Jeb Bush collection as the test case (not the Creeley archive depicted in the PowerPoint presentation).

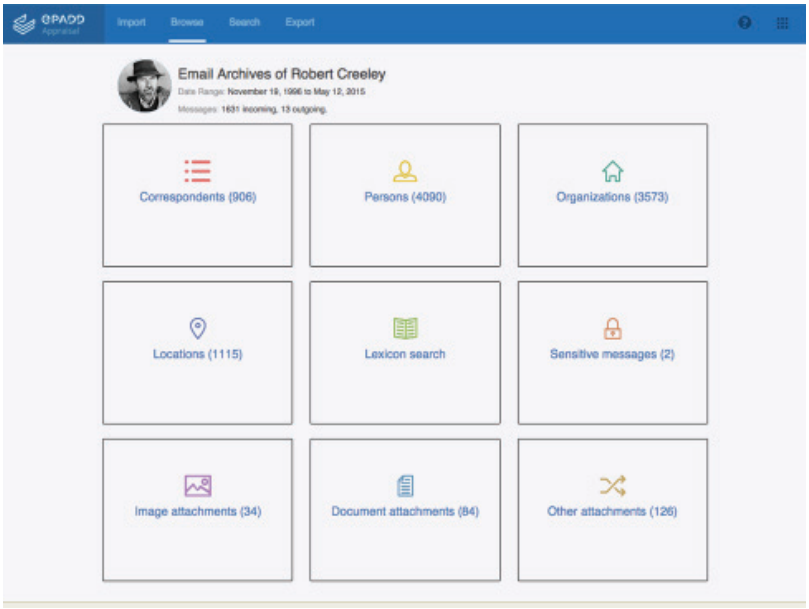


Figure 7 Browse Screen from ePADD PowerPoint presentation

She presented the browse page for the system, which appears in various modules in the ePadd workflow. One of the things she often goes to first when she's looking at an email archive, is the "Sensitive Messages" category. This is restricted data, e.g. SS number, credit card number, license numbers, etc.

The bulk of the demo focused on the Appraisal module. She focused on the Sensitive Messages view (Restricted Information) which included mostly constituent names, but also regular expression data, such as social security numbers, and credit numbers. Although the demo was of the Appraisal module, the processing module is very similar, but contains more powerful features, such as Authorities, that would allow an archivist to assign a particular name to a standard descriptive vocabulary or classification.

An important feature of both the Appraisal and Processing modules is that you can take actions against both individual and groups of emails. For example, you may choose to transfer, transfer with restrictions (and optionally, add an annotation), or flag as reviewed. This is useful as messages may show up in multiple searches. These icons (actions) do not appear in the online discovery module. In the delivery module several actions are possible; one may select the 'reviewed' icon or a 'cart' – to add a message (or group of messages) to your cart for export – or add an annotation. Exported messages retain the patron's annotations as an added header (see slide in question section).

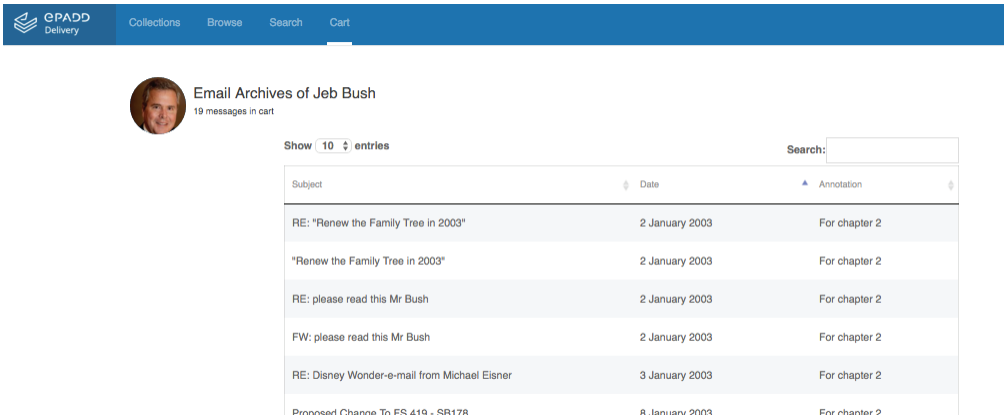


Figure 8 ePADD screen showing Jeb Bush archive (not in posted PowerPoint Presentation)

When bulk importing email into the appraisal module many bulk actions take place in the background, including deduplication of messages (e.g. in the Creeley email there were originally 150,000 messages, which were deduped to 50,000 unique messages), regular expressions searches (the regular expressions' lexicon may be edited), entity extraction (such as people, places, and corporate names), and name resolution for correspondents. The original default display shows results sorted by frequency - in either list or graphic format, but these results may be sorted alphabetically as well. While the algorithms attempt to resolve emails addresses into a single name, when variant addresses are found in corpus, the results may be edited manually (**Figure 9**).

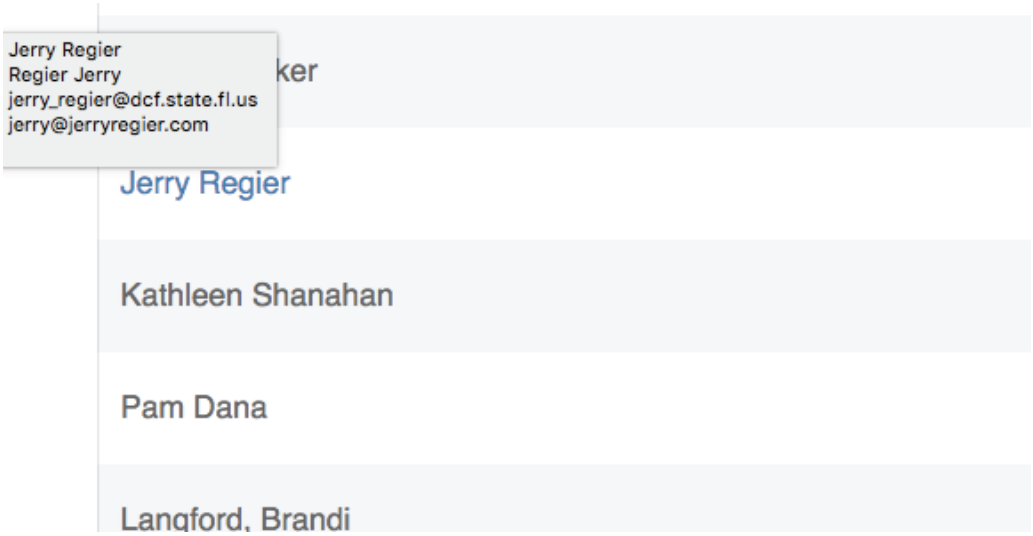


Figure 9 Name resolution – various email addresses resolved into one name.

Entities are also editable by the archivist during processing when selecting authorized versions of the names – ePADD is currently using OCLC's FAST which is included in the software. This is also incorporated into the program; once again, there is no need for an internet connection unless you want to view the DBPedia images when attempting to disambiguate between several individuals with the same name while performing authority work.

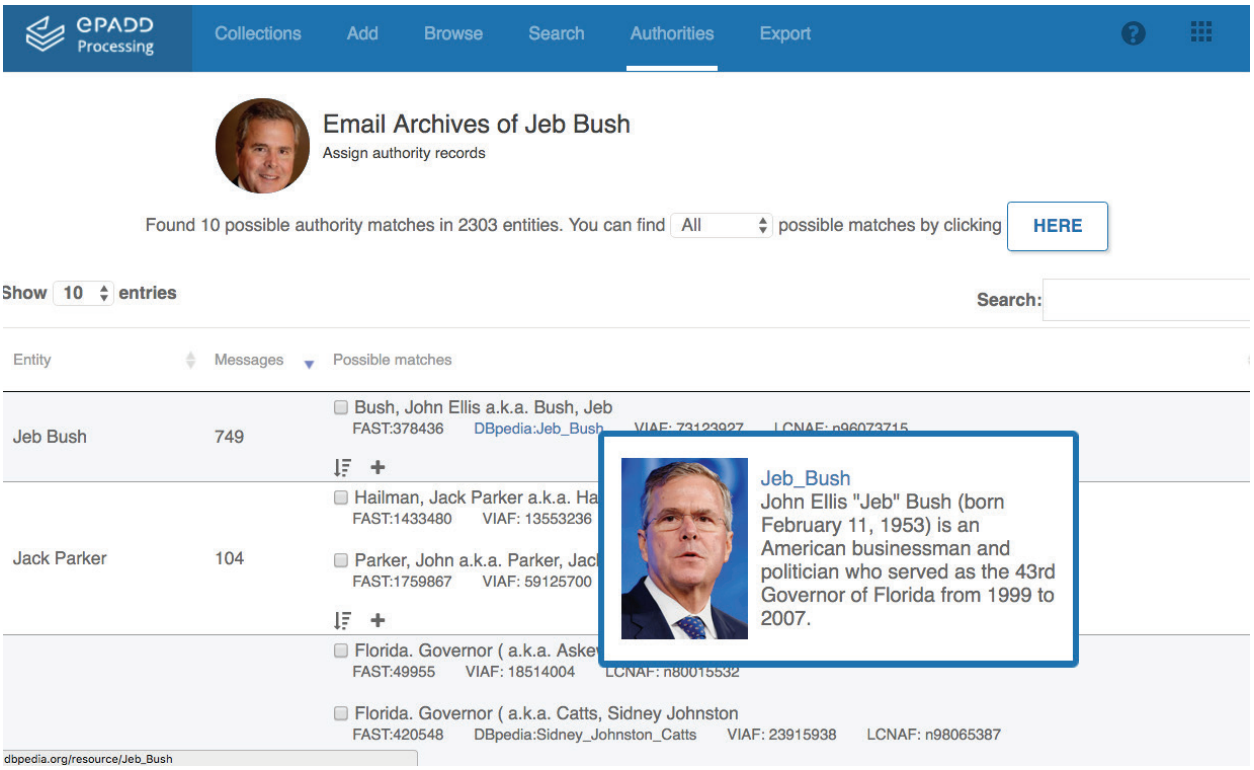


Figure 10 Confirming authorized version for Jeb Bush; can disable images by not connecting to internet.

If there is no correct authority – through FAST – you can create a local one in ePADD through a pop up window.

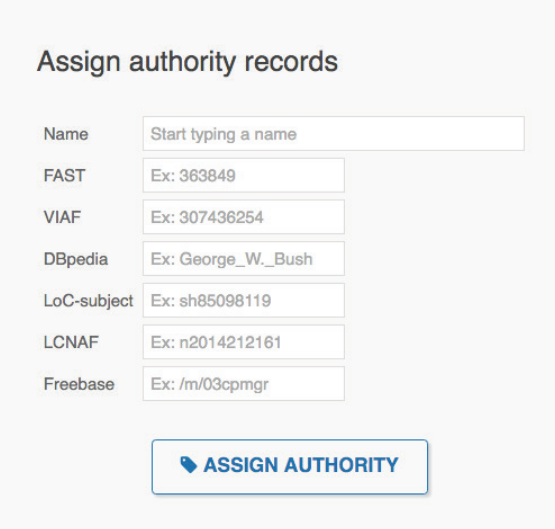


Figure 11 Assigning authorities in ePADD

Within an email archive, you can disambiguate first names that appear in the messages as well. ePADD provides a confidence ranking based on analysis of the corpus; the envelope means the suggested person is a correspondent as well.

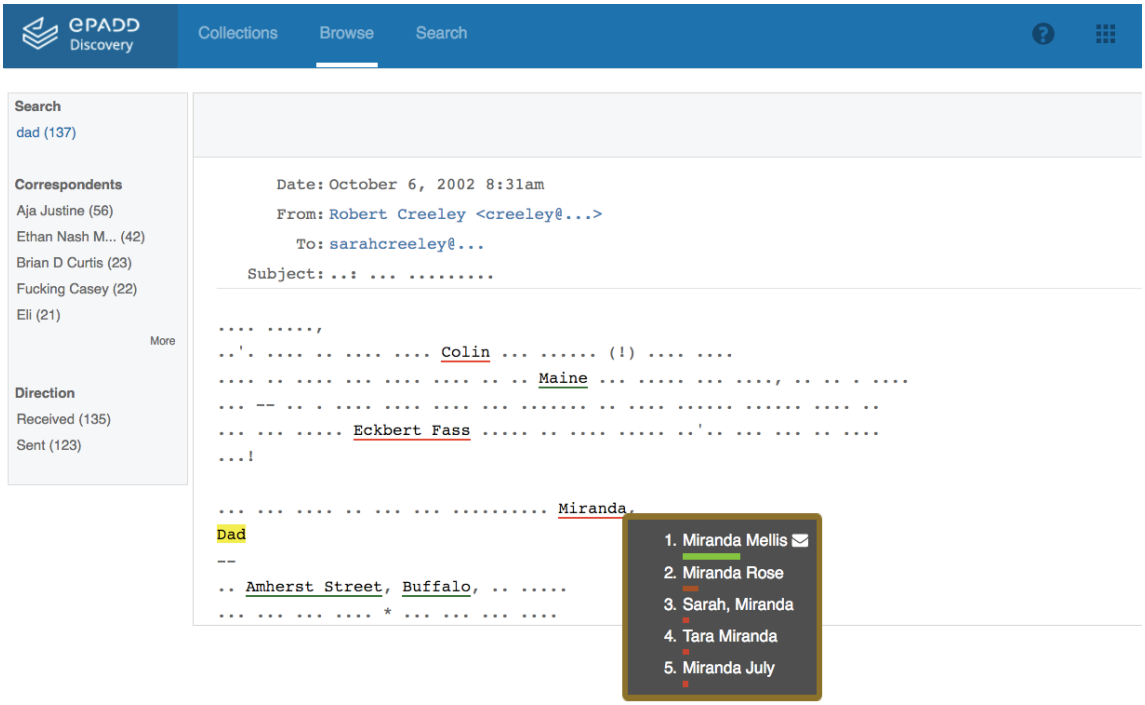


Figure 12 Disambiguating names within a corpus.

There was some Q&A regarding the tools used to do the regular expression searching. While some systems use bulk extractor, from the BitCurator suite, ePADD has its own method.

The image wall presents all of the images found in an email collection with links back to the original email messages. To get to the source message, you just click on the image (**Figure 13**).

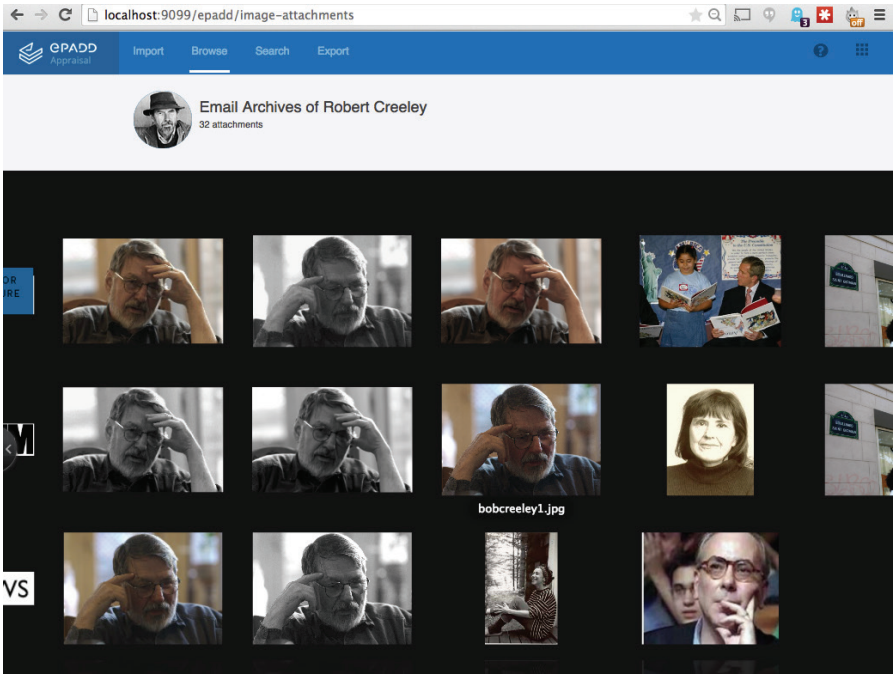


Figure 13 ePADD Image Wall

There are also several default lexicons included in the program that run preset searches – such as a "sentiment" lexicon and "sensitive" lexicon. The latter is meant to highlight messages that apply to employment, student records, and health issues. ePADD team believes that further development of better lexicons would be a great thing for institutions to collaborate on.

A new NER (named entity recognizer) algorithm in ePADD's current development cycle searches more extensively on several preselected categories – one such example is disease terms. For this new NER, ePADD selected main categories from DBPedia that we thought might be useful for researchers and archivists (e.g. "diseases and syndromes"). It searches all the main entries from DBPedia within these categories and searches for matching terms in the email archive. It's a fuzzy search: a perfect match ranks a score of 1, an implied result within the email archive will be ranked lower. The number of messages in which each term shows up are displayed as well. There was some discussion of the effectiveness of the DBPedia categories. The system allows you to click into the email to determine if there was something about it that might need to be restricted or handled in some special way. In general, it was far more effective than using the standard sensitive lexicon. Moreover, because terms for these categories in DPPedia are already loaded into ePADD, it is not necessary to have a live connection to the web to use this feature.

QUESTIONS FOR GLYNN

Questions were posed to Glynn during her presentation, and the following Q&A period. The questioning occurred in an informal, give and take manner, so in the following summary, it isn't always possible to determine who is asking and who is answering.

Q: Kari asked at what point would the tools not work as well as existing methods in an offline situation? She noted:

"How can you build-out those word lists and things that are local that you can then just upload at the time you're doing those searches? as another option? We build out these terms and such in order to be able to use them in an offline situation." – Kari

A: Glynn noted that one of the things they need to explore with other institutions, is what would they need to export for the system to be useful to other organizations, not just for preservation, but for metadata, for discovery.

Later enquiry by Glynn elicited the following answer from Peter Chan, ePADD Project Manager:

"NER with DBPedia and FAST, etc. are already in the setting files – no need to be online; we pull the images of personal entities from the Wiki live – so Internet access is needed."

More to the point of whether ePADD will function in a stand-alone configuration, Chan confirmed that if not connected to the Internet, ePADD just won't show images from the Wiki. All other functions will behave the same as with Internet connectivity.

Q: Kari asked what is the description that comes out of the system? MIT has not used the delivery portion of ePadd in their assessments, because it requires server installation and is more complex.

A: ePADD exports MBOX, but at present it is a complete environment, and does not produce an exported record for use in other environments. Currently, information is copied into other systems. However, exporting entities and authorized terms are part of the current grant cycle and will be explored in next 2.5 years. What Glynn had been referring to as "publishing" is approximately what others have referred to as delivery [e.g. into discovery and delivery modules.]

Q: Wendy asked if all the metadata added (annotations) to a collection of email appeared only in the header of the MBOX file.

A: Glynn followed-up with colleagues: Peter Chan indicated that metadata added by the archivist is stored in a mix of the header and separate files. *Sudheendra Hangal: "Re: the annotation, they can be exported in the mbox export and will appear as Xheaders."*

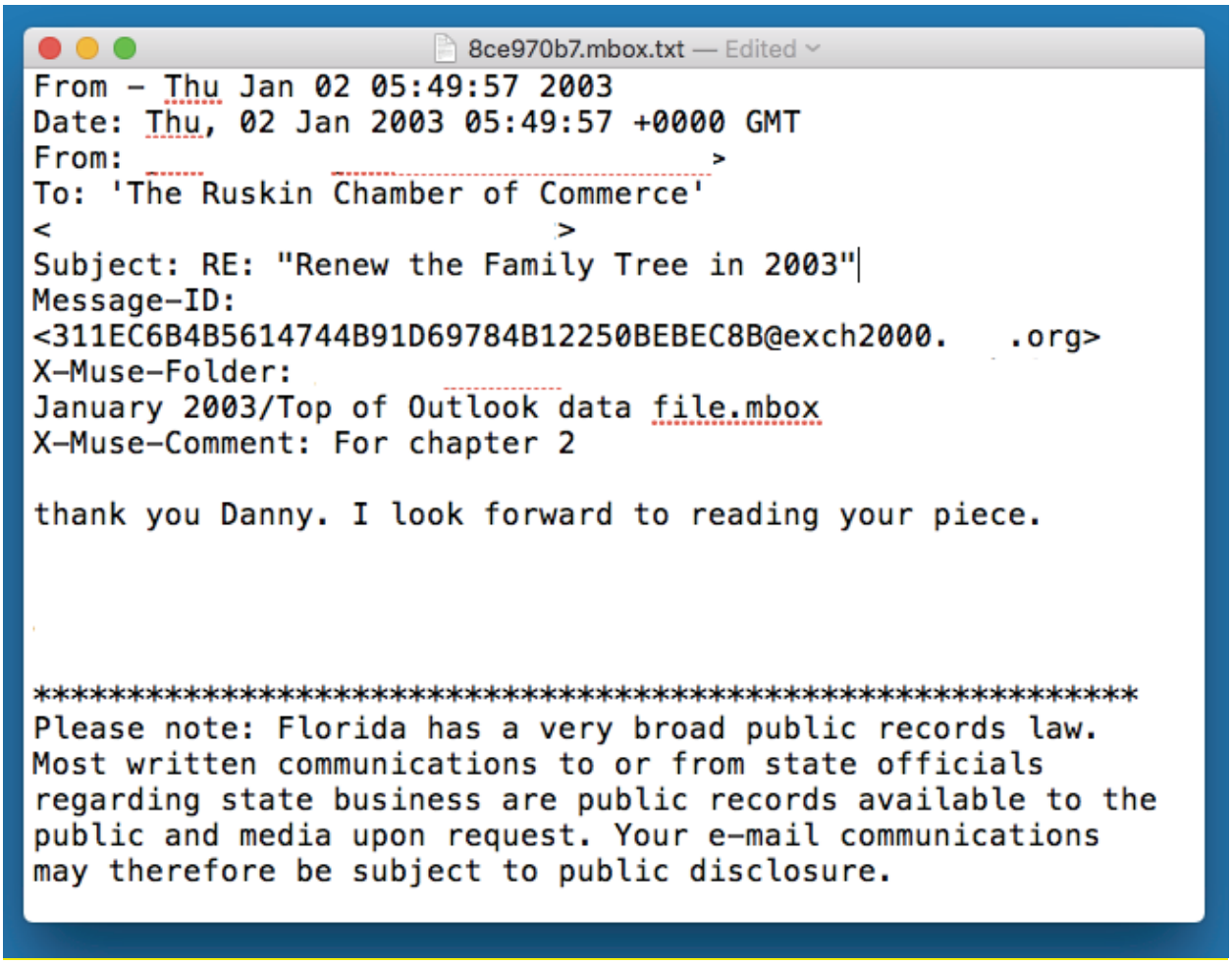


Figure 14 Sample email viewed in ePADD

In answer to a question about the creation of Marc and EAD records for the email collections, Glynn noted that all of their email collections (at this time) are part of larger archival collections, for which there would be a MARC record that referred either to the email that is available online (or other digital materials) or instructions on how to get access to these materials. The MARC record is a collection-level description. In EAD, they only have series level description, with some summary information. There followed a discussion of the location of metadata for digital materials.

Stanford, in fact, does not have a stand-alone finding aids site online for archival collections. Instead they use the statewide repository – the Online Archive of California.

RICCARDO FERRANTE — DARCMAIL

DArcMail is the successor to the CERP Parser (Collaborative Electronic Records Project.) The new system does more than CERP, taking advantage of MySQL and Python. The first email account they preserved at

SI was in 2005, and in the period 2005-2008, the largest account they preserved had 80K emails. Scale is a major issue for SI, and sizes have increased exponentially over the past 11 years.

The largest account they currently have at the institution is 30GB and the most recently acquired large account is 20GB, which suggests that this size has become the standard.

Currently their primary processing/acquisition tool is DArcMail.

One reason for the change in counting emails to counting raw size of an account is the variation in the size of individual messages, depending on their attachment sizes. Exhibition design staff tend to have "fat" attachments, while administrative staff has comparatively "skinny" attachments.

He presented a modified version of the "chart" that compares the abilities of different systems to perform the activities that make up email archiving (EAST2016_DArcMail_Demo.ppt, *Slide 3, http://bit.ly/1T9VGLw*, **Figure 15**). He pointed out that the way the chart is laid out suggests a certain flow, but the way the tools work doesn't necessarily follow the chart. The green areas on the chart he displayed represented the things that they're using DArcMail to accomplish.

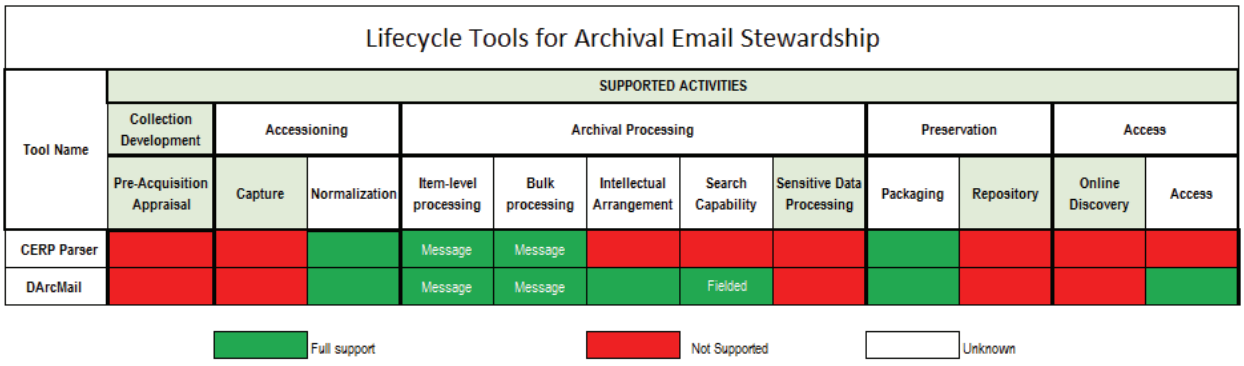


Figure 15 DArcMail presentation, from Slide 3, modified chart

The CERP parser could work on one message or a whole account. If it were a really *really* large account they would set it up on a machine and let the computer run until the job finished, which might take a

whole day - unless it bombed along the way. It converted MBOX to the Email Account XML Schema, developed in collaboration with the University of North Carolina. The schema captures all of the email components, and references all of the attachments, so the xml file itself can represent the whole account, but not the calendars and journals. It includes the whole folder structure that's there, all the way through any kind of threads going back and forth. The CERP processor would then create a subject sender log, and some additional harvested metadata from that. It would package the original source of the account in an MBOX, which was kind of their interim normalization, and then create the xml, which is their preservation master, and the other files. All the components were open source; the whole environment could be set up on a flash drive.

However, Squeak is not a very popular open source platform nor is its parent, AppleTalk. Raw xml is ugly. They got as far as creating the results but they didn't have a GUI for it, and GUI is how people work.

So eventually they created *DArcMail* and added in things they couldn't get to previously for lack of funding.

An important thing to keep in mind when considering how DArcMail is designed is the context of the archives in which it was created. DArcMail works very well for the SI Archive's workflow. Some aspects of that workflow:

- Appraisal in general is a precondition to acquisition.
- Most of the email they've taken in is part of personal papers, or the correspondence of particular people at the institution, so it's paper plus. They have very few accessions that are digital only, unless they're websites or social media.
- The documentation of their collections happens in their Collections Management System.
- Their digital preservation steps are as preemptive as possible, which means that as soon as digital material comes in, their electronic records archivist is pulled in.
- They review and do risk assessments on everything.
- They work with the acquisition archivist to define the scope of what should be accessioned and they try to avoid deaccession, but it also means they do bit-level preservation as things come in.

Nonetheless they're trying to spend their money to make tools that other people can use. Goal with DArcMail was to allow archivists to understand the account, and still do preservation. He compared CERP to DArcMail:

The CERP parser just preserves. DArcMail includes searching and exporting, simple GUI, is 4x faster than CERP, and uses Python and MySQL, which means it can be used in a standalone setup. It can also run in a client server configuration. It runs on a variety of platforms.

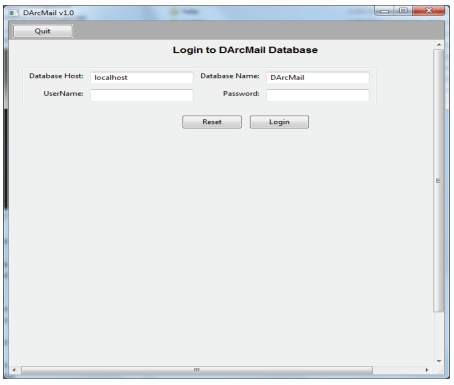


Figure 16 DArcmail presentation, Slide 7: Logging in to DArcMail

Slide 8 (**Figure 17**): The user defines an account. The account, as it's pulled into the database could be on a variety of different work locations, such as a workstation, a shared drive — anything so long as you can get to it. It can process just the top level folder, or work its way all the way down the directory tree. It works off of MBOX.

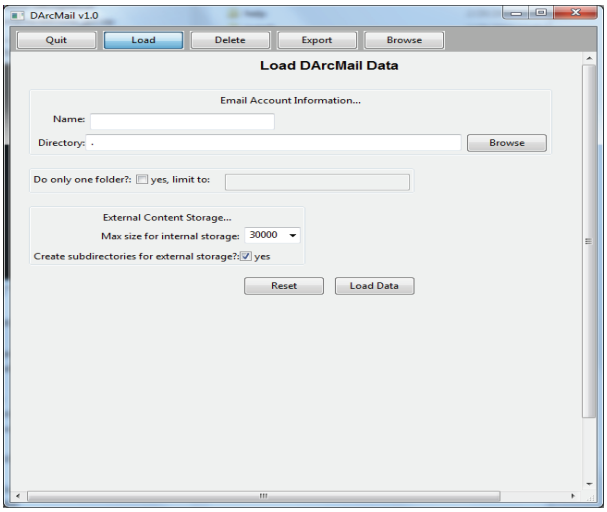


Figure 17 DArcMail presentation, Slide 8, Defining and account

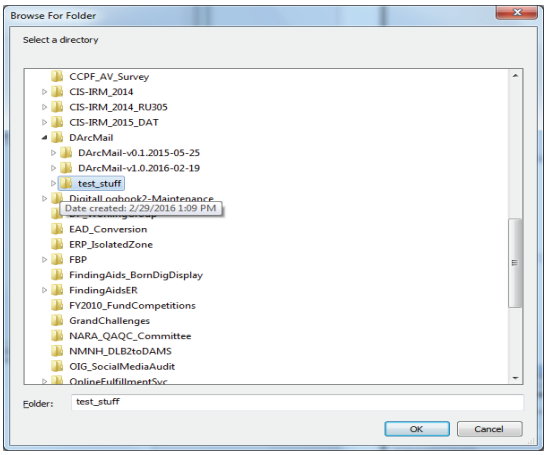


Figure 18 DArcMail presentation, Slide 9, select material from source location

Workflow: They can load the account from almost anywhere, and the system provides feedback on how successful the load is. They also have the option to delete the account from the db. A motivating idea for this function is that they can load the system as a standalone in the reading room for a researcher with all the accounts that they are authorized to view, and delete the rest from that installation.

They have the ability to browse. To do this, you set your account. The search frame helps you identify which one it is. He chooses Clough. You can search fielded keywords on the messages. He gave the example of searching "newsletter" and "Wayne". You can limit search to plain text, HTML or both. Thus, they can search the plain text in the message itself, and also metadata associated with the messages. He

has sort order with a variety fields. He can search on address name, but for email that doesn't go outside the system they don't record the full email address.

Attachments stay in their original format. The system allows you to download any plugins necessary to use an attachment. When a message is retrieved and displayed, the search terms are highlighted in the body text of the message. If you have an attachment that is used in a number of messages, you can search on part of the attachment's name and the system will bring the associated emails back.

The search capability is not true full text, or true advanced Boolean. In terms of access, they only have workstations in their reading room.

QUESTIONS FOR RICC

Q: Is Reading Room access mediated in a manual way, or are they able to load just the things a researcher wants to see and step out of the way?

A: When a researcher is approved, a fully working setup with only the material they're authorized for would be installed, and after appropriate training, they would be left on their own. DArcMail is used to export a version of the content that has been tailored through the above mentioned deletion capability.

Q: Are there facilities for automatically redacting sensitive information?

A: They do not do any redaction. If someone asks for something, someone has to go through it. They don't use regular expressions or automated clean-up.

Q: What kinds of export formats does the system provide?

A: For export they can push all content and Metadata out as XML, but in the reading room context, the researcher is interacting with the accounts in *MBOX* format.

It will be on Github when Ricc finishes the user and installation guides, but several panelists expressed interest in getting the source code. They do have a PowerPoint that talks about table structure etc.

There were questions about the formats and conversions necessary to make things work in other systems being presented. The question arose as to how much work had been done to convert email formats to *MBOX* (which seemed to be the standard format that developers around the table were using) particularly for older formats, like Pine.

SKIP KENDALL — ELECTRONIC ARCHIVING SYSTEM (EAS)

Skip presented a live demo of the EAS system, which highlighted the new EASi interface to the system

The system is ultimately intended to handle all electronic content, but currently is designed only for email.

There is no public user interface at this point, only an administrative UI for curatorial ingest, processing, and depositing to the preservation repository.

Whereas Stanford is working to develop cross collection searching, in EAS this is the default. When you first open up the system, you are presented with everything in the system within your account. Results are brought back from a search across all collections, and selections from these results may be made. You can choose one, or multiples up to everything initially retrieved, and perform actions on them. This includes assigning tags at the collection level, assigning items to series, assigning processing levels, and billing codes.

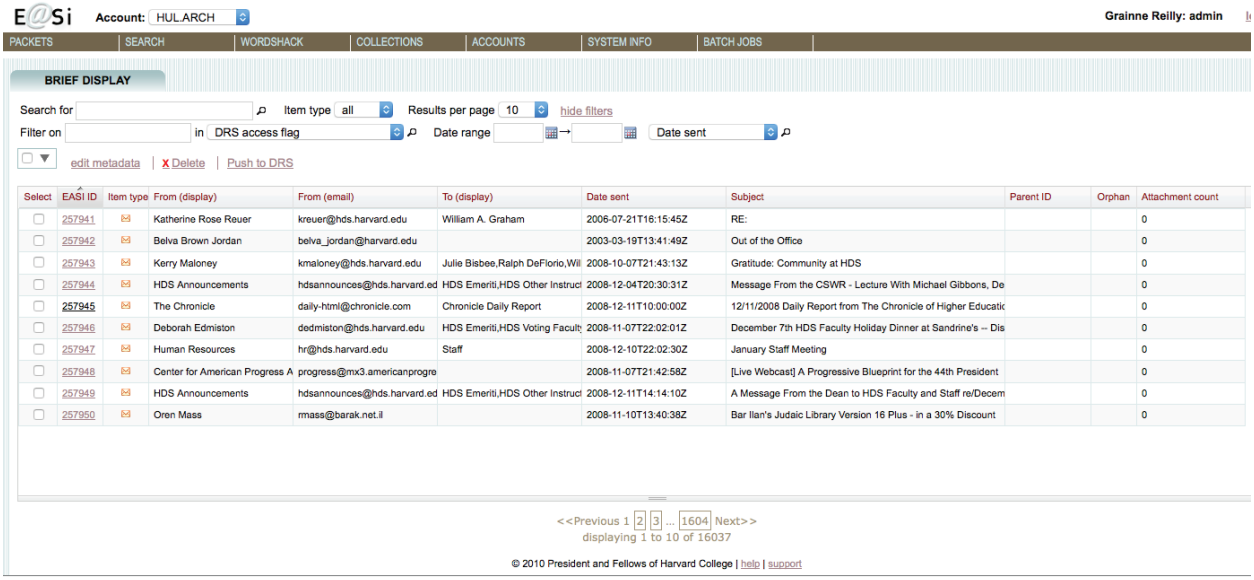


Figure 19 EASi Brief Display

They can also put notes onto objects, including public and non-public notes.

They would like to be able to process things immediately, whether at the item or higher level, but in practice, the volume of material makes this impractical, and their practice is to assign some minimal level of metadata to the incoming material, and then send it into the preservation repository, where it can be more fully processed later.

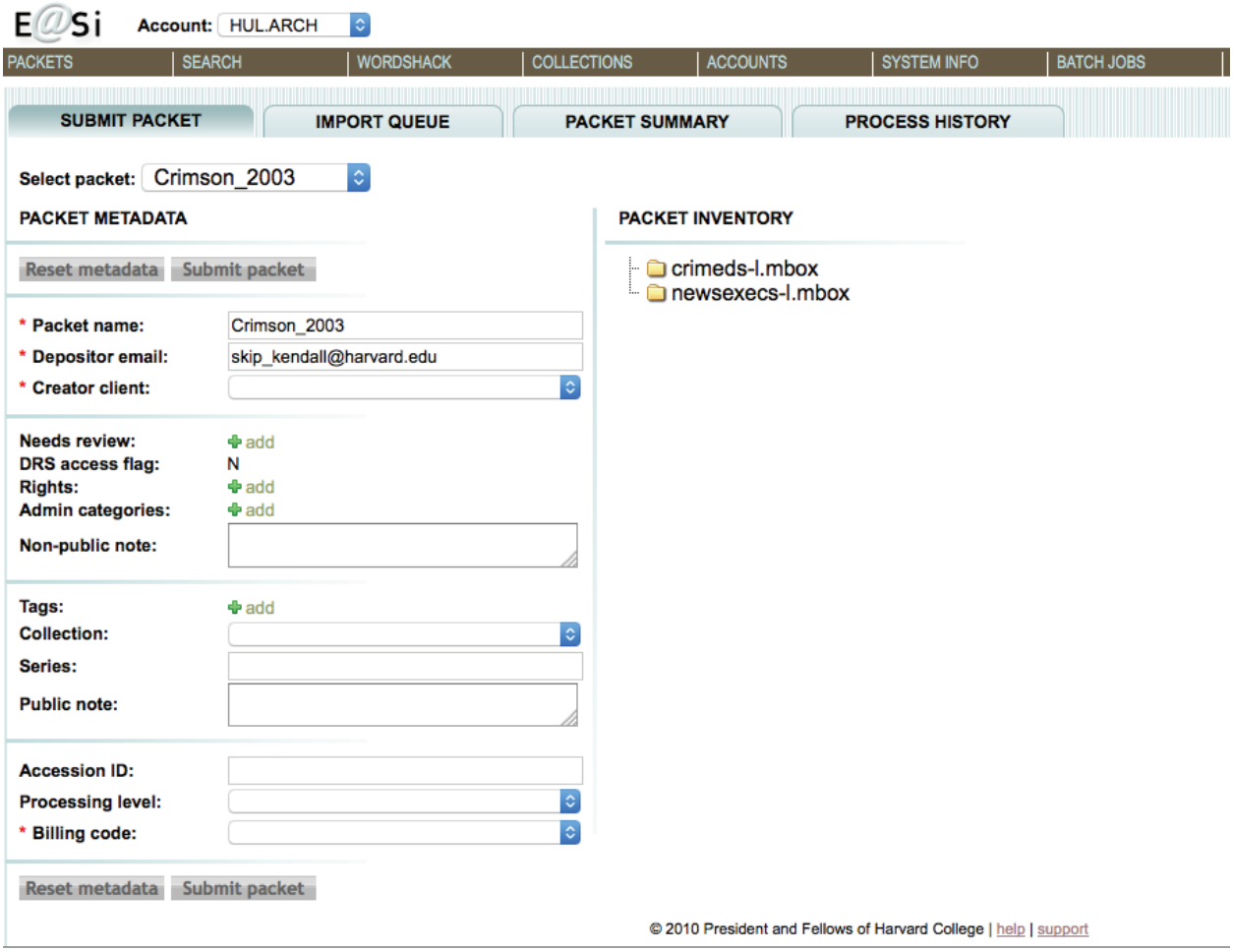


Figure 20 EASi Packet Submission

Regular expression filters are run during the ingest process and generate flags. During processing, archivists can isolate objects with these flags, review the flagged content, and remove the flags. At the point of removal, appropriate action (such as deletion of the object or application of restrictions) can be taken. The flags can also be left in place to review at a later date.

The system includes Digital Repository System (Harvard's preservation repository, hereafter referred to by its initials as the "DRS") access flags -- 3 levels of these:

1. Not accessible (except to authorized account holders)
2. Open to the public
3. Restricted access to Harvard only.

But currently there is no delivery system! (Only authorized account holders can retrieve email and messages and therefore need to mediate use by end users)

There is a section for Rights information.

There is a section for Admin categories, which can be used for non-standard group creations. An example was given, indicating a set of materials that were used for an exhibition.

They can choose from 3 levels of processing:

1. Completely processed
2. Not Processed
3. Partially processed

The processing flags can be set for individual messages as well as the entire collection. So they can mark one part done, and another part not done.

Needs Review. This is a note flag that allows a processor to alert a supervisor or other cataloging staff about issues that need some decision to be made. Processor will begin the note with the supervisor's initials. Supervisors can periodically search for their initials and find items that require attention. This search is applied as a filter, and will retrieve only those items marked. The flag can also have a review date associated with it.

Search filters can be removed one at a time, allowing for incremental processing.

There are currently no individual permission levels established for what people can see or do. The permission is currently set at the DRS owner code level. An Owner Code is the coded representation of a digital object owner (a Harvard organizational entity with financial and curatorial responsibility for objects in the DRS).

Although they can add metadata at the item level in batch to large groups of records, they do not have the ability to add metadata at the folder level, as there is no explicit concept of folder tree in the system. The folder tree can be virtually extracted, since all of the records include a path.

Participants agreed that being able to manage email by folder structure would be useful, rather than assigning it to individual items by grouping on a path, as it would make the creation of series easier. Skip pointed out again that what they were seeing on screen was everything in the system, and not just one account.

The system uses a tabbed interface that allows the processor to drill down in a left to right direction, by clicking on items in a tab. The message tab shows the actual email message with a metadata panel on the left side. TO and CC fields are there, but the message did not have a display name for the CC field, so it remains blank.

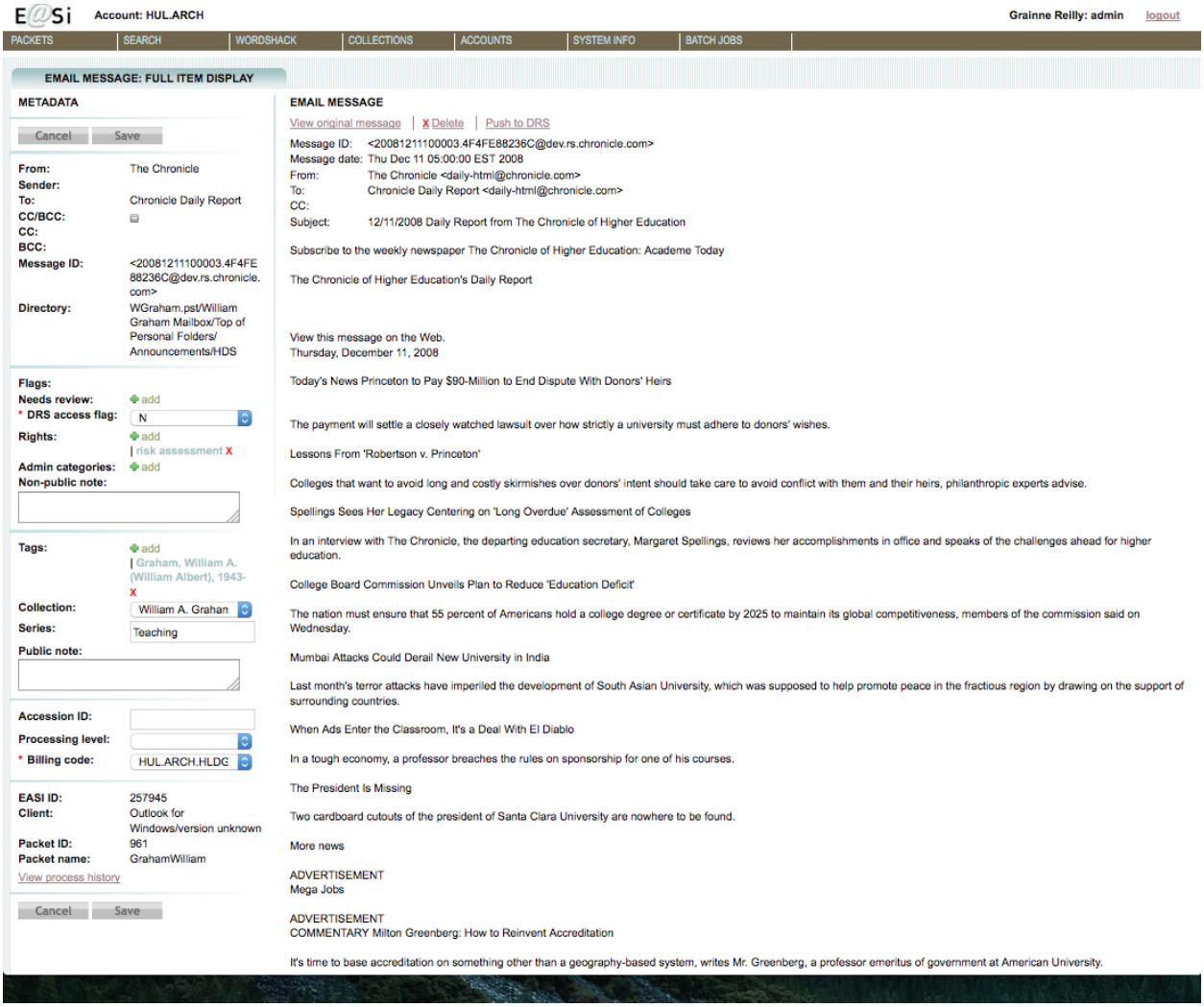


Figure 21 EASi Mail Message Screen

There is a risk assessment assigned to messages, and at present all item receive an "secure storage - unconfirmed" tag. All of the material goes into a secure section of the repository.

There is a process history where events such as conversion to EML from the native format, and deletes are recorded (**Figure 22**).

On ingest, they use EMailchemy to convert to an EML file, then extract embedded attachments, and rewrite line breaks.

They also record deletions in a PREMIS record.

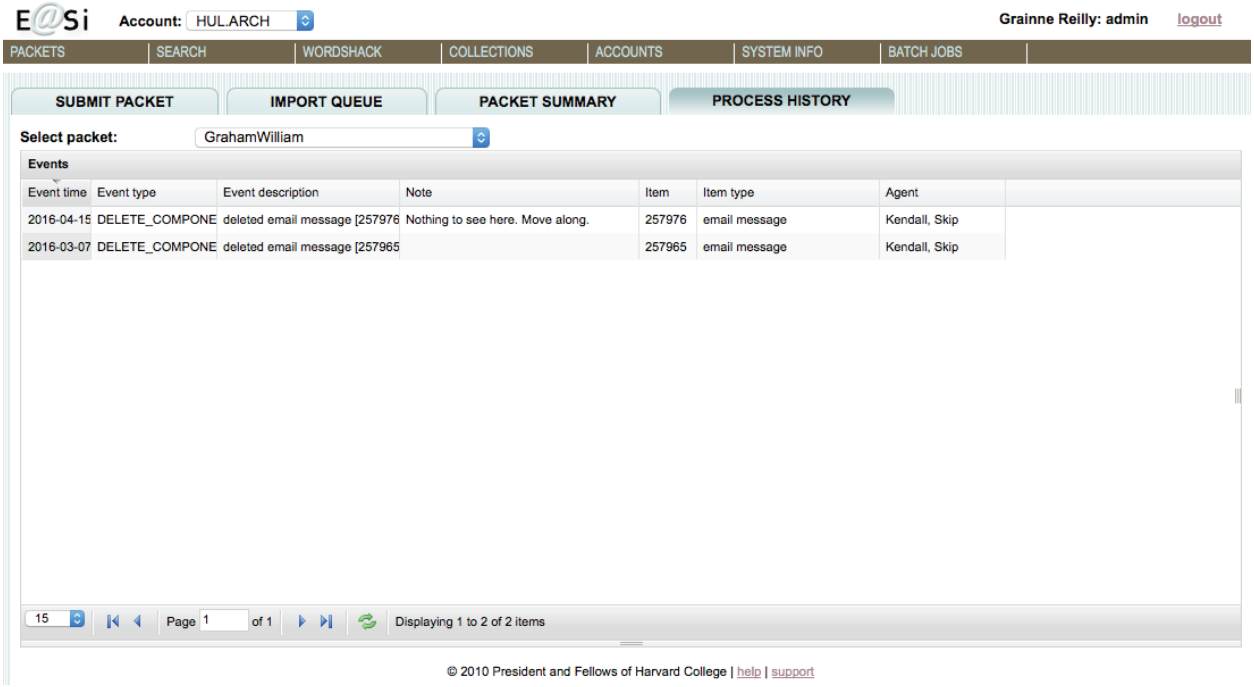


Figure 22 EASi Process History Screen

There are links to and from attachments so that you can get to and from attachments wherever you find them.

Although under most systems throughout Harvard, pushing to the DRS is complex, this is one of the few systems that has built-in routines allowing it to be done with a simple click. Individual items or groups can be pushed to the repository.

DISCUSSION

There was no formal question and answer period for this presentation. Instead an open-ended discussion of the system functionality took place, during which the following points were made.

Wordshack is an application that was developed in parallel with EAS, and is used to associate names/email addresses. A discussion took place during this portion of the demo concerning the ability to resolve names and addresses on email that is being ingested. Skip was asked if there were links in the system to the Harvard directory to positively identify names in the email send and received fields, since internal mail does not include the full email address. Currently the system does not do this, and archivists must manually do this resolution. Chris mentioned seeing a commercial product in development that does this kind of name unionization.

Cal Lee was interested to find that the system could generate PREMIS data that could be pushed to the DRS, rather than being created at the time of ingest into the DRS.

Justin Simpson noted that they are working on a feature that incorporates PREMIS data from other "agents." This is included in the METS file. They are looking at a Fedora implementation to do this, but this is something that their clients (including the Harvard Business School) would like to do.

The system is dependent on the concept of "packets" which provoked some discussion. Packets are groupings of source emails, and associated metadata that is treated as a discrete unit for handling the material. Although described as an arbitrary unit, it maps closely to the concept of "accession," in that it's a way of grouping material that moves through the workflow as a sub-dividable unit. Packets can be

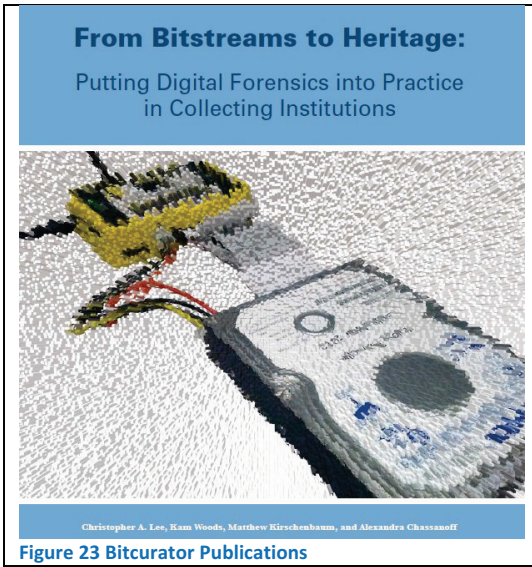
grouped into larger collections or subdivided when fully processed. One of their important features is that they can acquire their own metadata, so for example, if an email is deleted, the metadata indicating the deletion would be attached to the packet object. This “packet” concept is also found in Archivematica where it is called a "transfer."

CAL LEE — BITCURATOR ACCESS ENVIRONMENT

There is a pdf presentation: harvard-east-bitcurator-demo-20160302.pdf (http://bit.ly/1LNBXzI)

The goal is to develop a system for people in the LAM community to incorporate the functionality of open source tools for various forensics tasks, though they are concerned about calling it "forensics."

BitCurator tries to incorporate these functions into the workflow of archives and libraries. In their publications (**Figure 23**) they tried to build a community around these tools that could be used for the archival community.



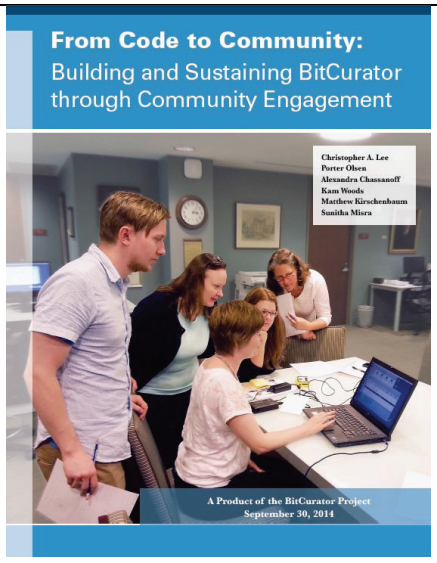


Figure 23 Bitcurator Publications

(Publications available at http://www.bitcurator.net/docs/bitstreams-to-heritage.pdf, and http://www.bitcurator.net/wp-content/uploads/2014/11/code-to-community.pdf)

BitCurator runs as a Linux operating system, that combines a bunch of tools into a single environment. It has particular applications, e.g. FITS, that are built into it that will be useful for various groups. And they are willing to incorporate other tools that are brought to their attention. There are tweaks to the standard Ubuntu Linux environment to make it friendlier for the work, such as defaulting mounted devices to a read-only configuration. There are also environmental menus that come up on a right-click. It is not a workflow based suite, like Archivematica, but compliments that system well. It's more of a platform for the initial stages of acquisition where you piece together what it is that you need to do.

It can be run as a virtual machine, but the best way to run it is as a true Linux environment.

The system is maintained by the BitCurator consortium (https://bitcuratorconsortium.org/).

Although redaction is listed as a function on slide 12 **(Figure 24)**, BitCurator does identification of items for redaction, but does not currently do actual redaction. Redaction as a function is something that the consortium is contracting for in the second year of their grant. There is a Python library for doing redactions, but it's rudimentary.

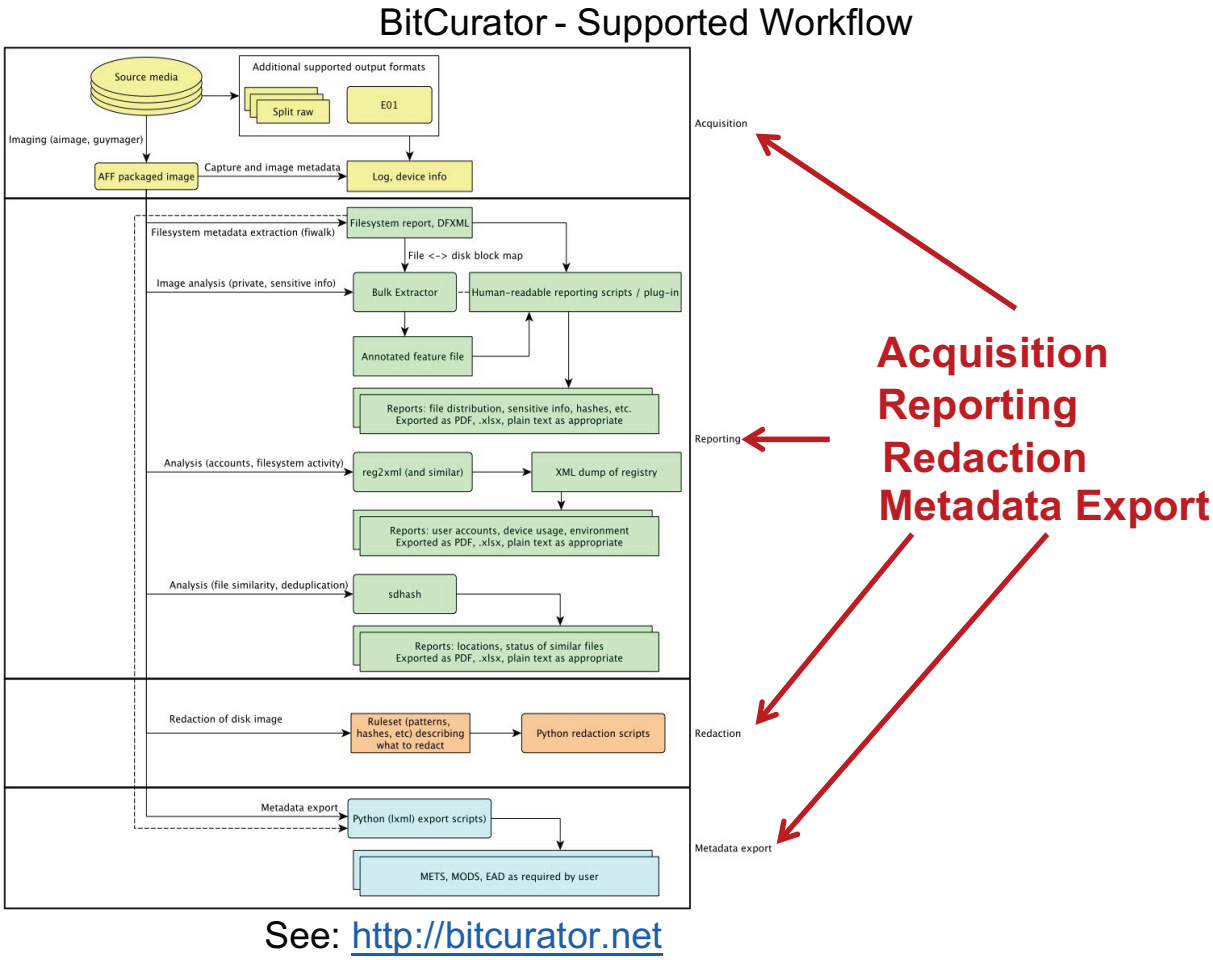


Figure 24, BitCurator presentation, Slide 12

The core of the system is DFXML, which has information about the file system; it includes information about the directory structure, time stamps, user accounts, sizes, et.

Slide 13 **(Figure 25)**: Disk Image. You can choose in the environment how you want to get the image. You can choose a raw image, or packaged one that includes metadata. You can assign metadata and send it off with the raw data to create a disk image.

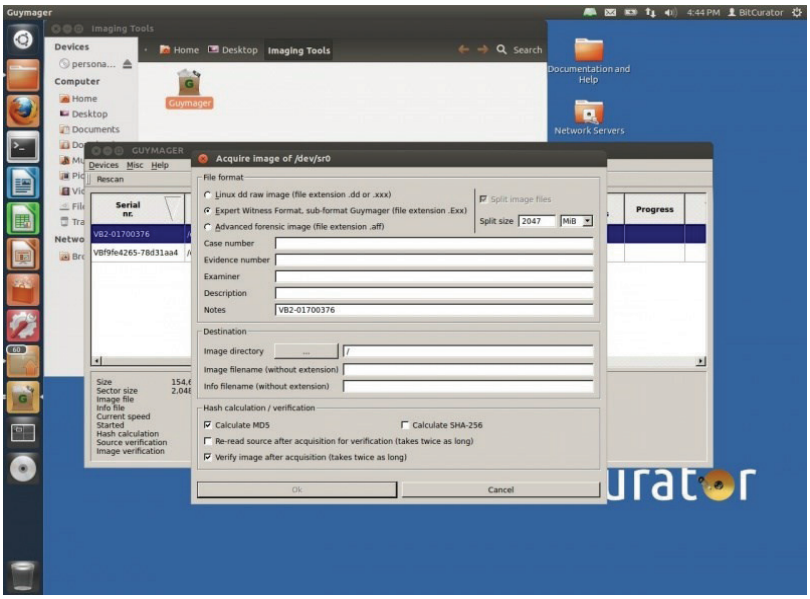


Figure 25, BitCurator presentation, Slide 13

For example, there is a script that allows you to right click on a disk image and choose an option that will retrieve information about the disk image. Where there is great complimentary functionality between

forensics and archival work is in the area of provenance. All of the information about creation and modification is embedded in the BitCurator tools.

The fundamental unit in BitCurator is the disc image, so you are always starting with that, thus allowing you to capture the context of email (if that's what you're looking into) – essentially it allows you to establish original order in the digital context.

Ways to interact with disk image:

- Mount them as regular drives, allowing you to navigate through them as with any other drive.
- Another way is to inspect them using the forensic tools, using the Disk Image Access tool, or the new VCA web tools, that allow this to be done over a web browser.

Disk image information includes the Operating System, file information, including directory entries for deleted files, thus allowing you to make decisions about exposing this data to others, including processing staff. You can choose to only export the allocated files, and not hidden or deleted files, to a new disk image.

Bulk Extractor. Allows you to identify information that might be sensitive. They continue to add scanners

Slide 20 **(Figure 26)** shows what Cal calls a histogram of an email.

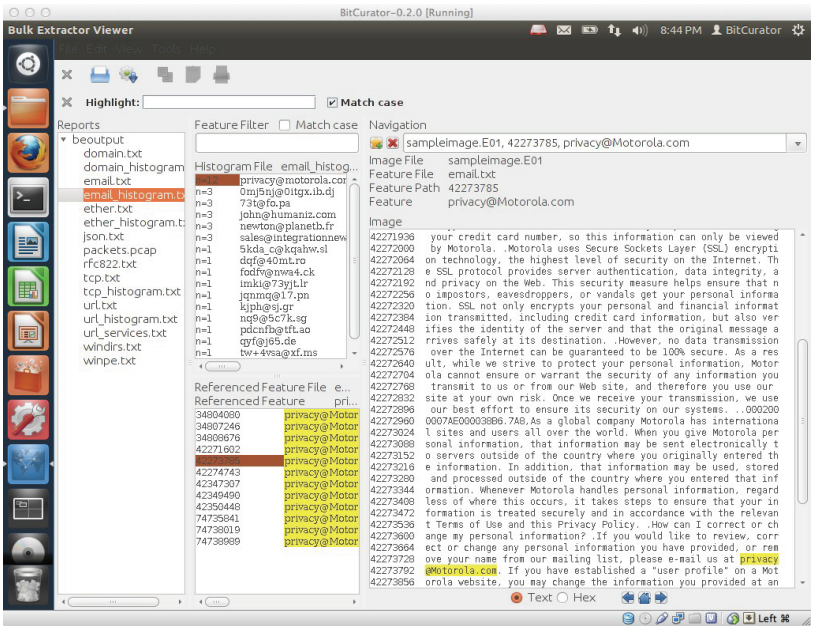


Figure 26 BitCurator presentation, from Slide 20, an email "histogram"

Slide 21 **(Figure 27)** includes sensitive information (blacked out) from the Jeb Bush archive that had been posted online briefly. He points out that though it's difficult to locate certain kinds of sensitive information, there is a great deal of information that is findable through patterns, and that searching for these with a tool like bulk extractor (http://www.forensicswiki.org/wiki/Bulk_extractor), is relatively straight-forward.

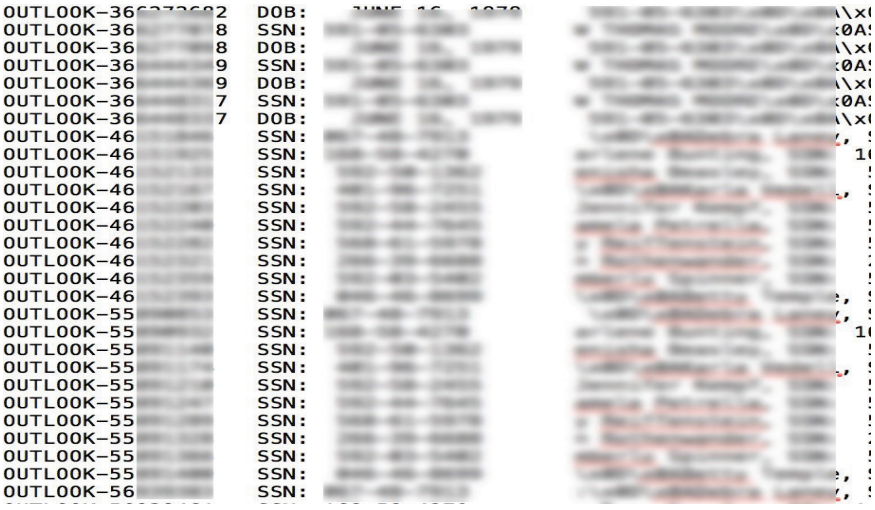


Figure 27 BitCurator presentation, from Slide 21, sensitive information (blacked out) from the Jeb Bush archive. SSNs and DOBs identified in large PST collection using bulk_extractor

Cal ran through a demo of fiwalk XML which produces a DFXML file. DFXML is a schema maintained by NIST, available on Github, which is used to describe a disk image. This file contains provenance and technical metadata, stuff about what the software was, how it ran, what commands were issued, low level information about the disk, such as how many sectors it had. Information on "file objects," which describe every directory or file on that disk, including remnants of files that were in the directory, but were deleted, are described by the xml. They also are able to export the same information as an Excel file, because many people find this easier to work with. Also, other applications in the suite use Json, which cannot use the xml code, and need the data in another format such as Excel. Nonetheless the schema entities and relationships are present in whatever format is used.

DFxml is the common export format used by open source digital forensics tools.

In addition, you can always go to the command prompt and execute commands for the tools that cannot be done through the GUI environment.

Another tool that he covered is READPST, which is used to export data from a pst file to MBOX format. The tool allows you to export entire accounts, individual messages, or folders in the pst, and have them reflected as actual folders in the MBOX. There are many switches available to do this. READPST appears to be run from the command prompt.

BitCurator Access (BCA). This is a new project to create an interface to allow end-users access to the things that have been acquired with BitCurator tools.

Slide 38 (**Figure 28**) - Automated redaction and Access options. This relies on "emulation as a service." Cal describes this as a platform that allows you to feed data to it, then it emulates the original display platform and distributes it via the Internet and is viewable in a web browser? Material that has been redacted will not appear.

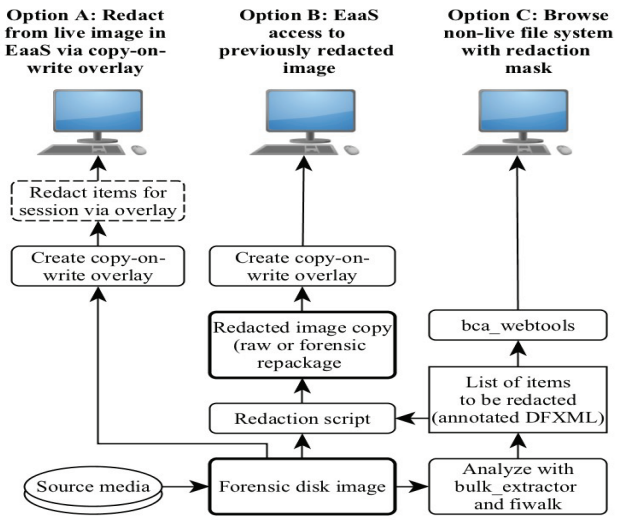


Figure 28, BitCurator presentation, from Slide 38, Automated Redaction and Access Options

Slide 40 (Figure 29): BCA is a service that sits between the BitCurator tools running on a server, and allows access to their functionality via a web browser, and an Internet connection. Most of the analysis runs on the server. Three scenarios in which this setup is useful:

1. Internal. This is a convenient way to navigate around the disk image as you're making decisions and processing the material.
2. In the reading room, where it can be used as a lightweight way to provide access to the content.
3. The open web, where people can search and navigate the content.

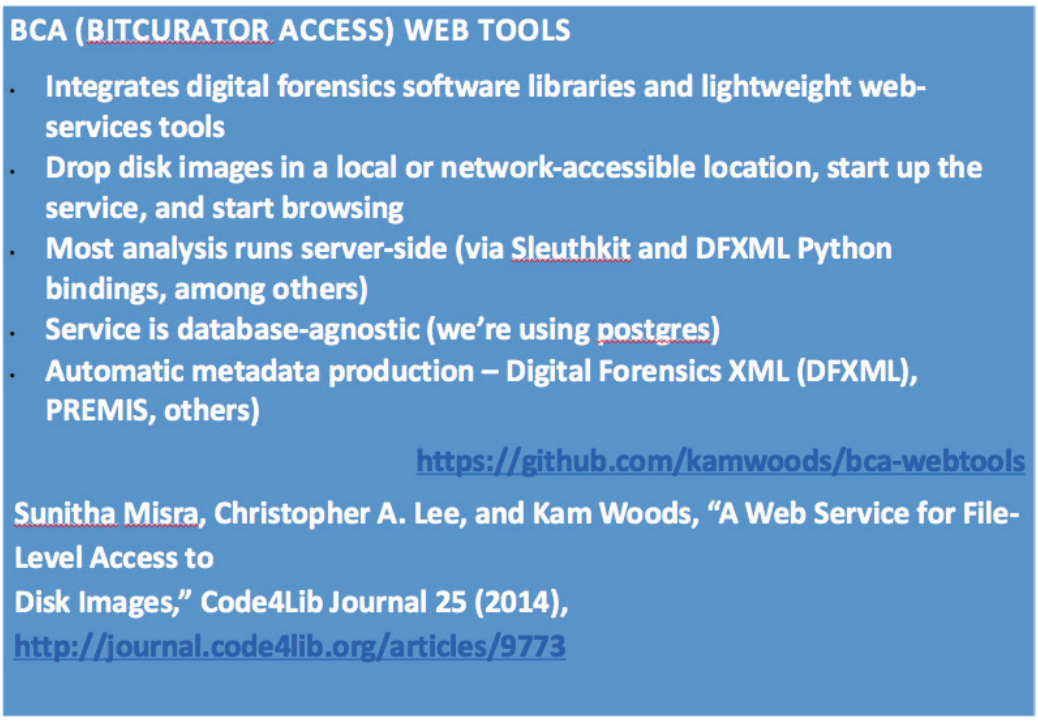


Figure 29 BitCurator presentation, from Slide 40, BCA Access Tools

Slides 41-42 (**Figure 30, Figure 31**): Shows a diagram of the BCA environment, which presents the BC apps running on top of a stack of web services and utilities that provide searching and indexing functions. This is conceived of as an end-user interface. The idea is that you "drop" disc images into it, with minimal processing. There are options to do further processing, but if you only drop the disc image into the system, a set of small xml files is produced representing the results.

You can choose objects from a directory listing and open them up from there just to look at them, and get a sense of what is in the material.

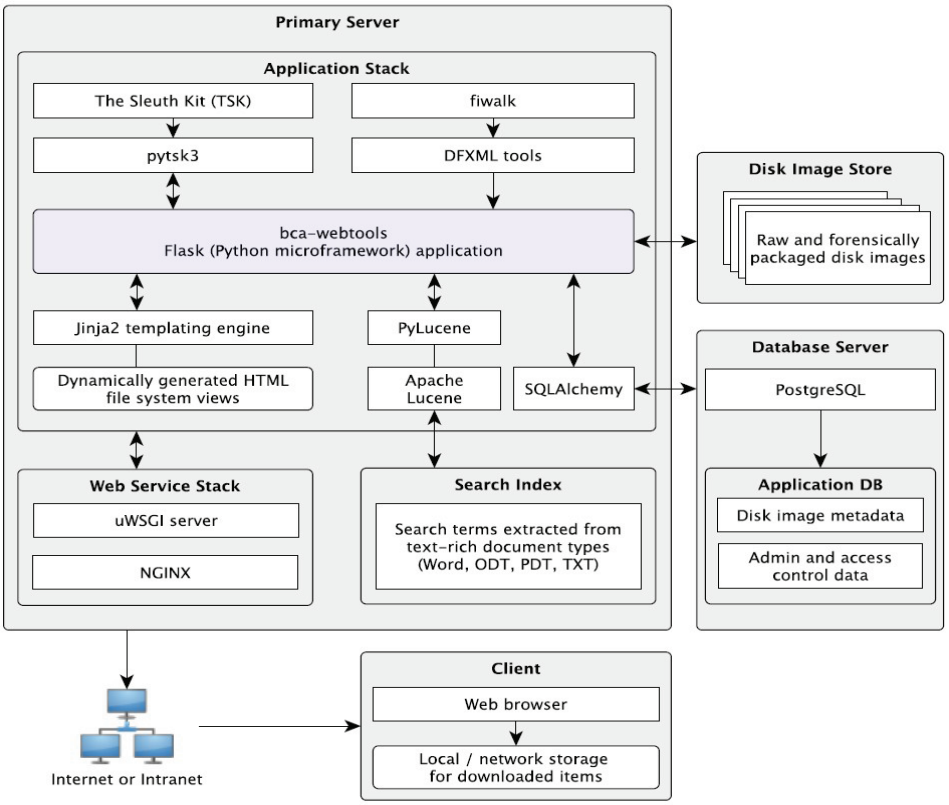


Figure 30 BitCurator presentation, from Slide 41

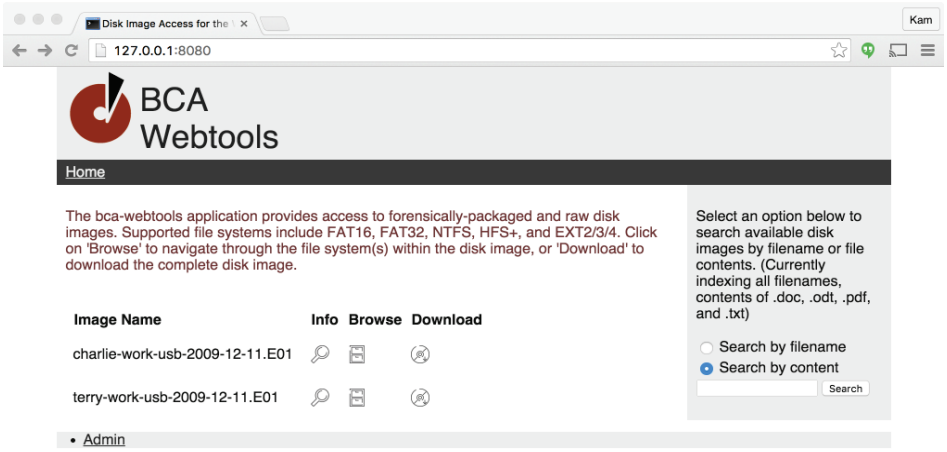


Figure 31 BitCurator presentation, Slide 41, BCA Webtools

Slide 43 (Figure 32): The admin screen provides information about the level of analysis that has been performed on the image. The image matrix indicates the existence of a DFXML database, whether the disc is indexed, and provides the option to create or delete these items. The indexes and DFXML DB's are necessary for search and retrieval of objects in the image, but depending on the size of the image, the production of these (click add, and then submit) could be time intensive.

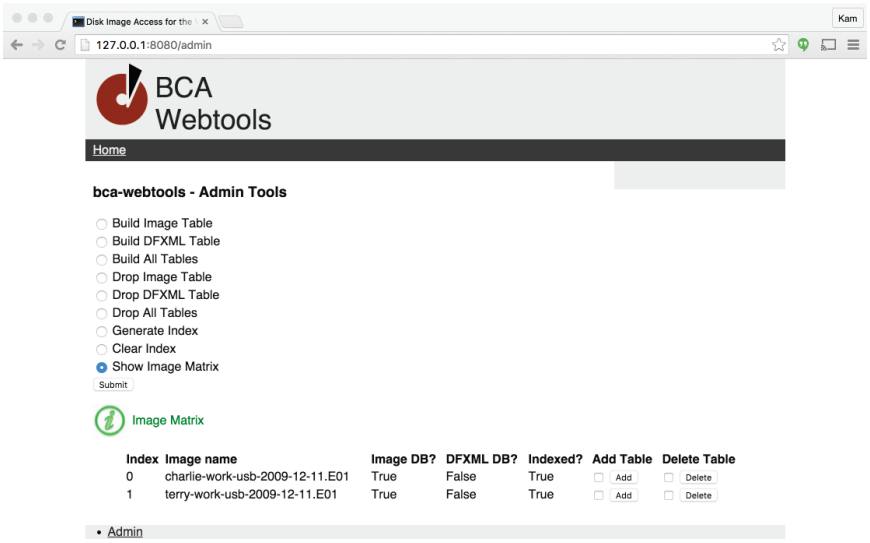


Figure 32 BitCurator presentation, Slide 43, BCA Webtools

Slide 44 (Figure 33): The end user interface. Click on an image, and the list of items are displayed. The list items are linked to the objects, so clicking on any of them will retrieve the object from the disc. Cal did not show the search screen in the system. But you can search both by filename or content. The system is very basic, but it was built with a RESTful API with the assumption that other software would be built on top of it to provide more sophisticated search.

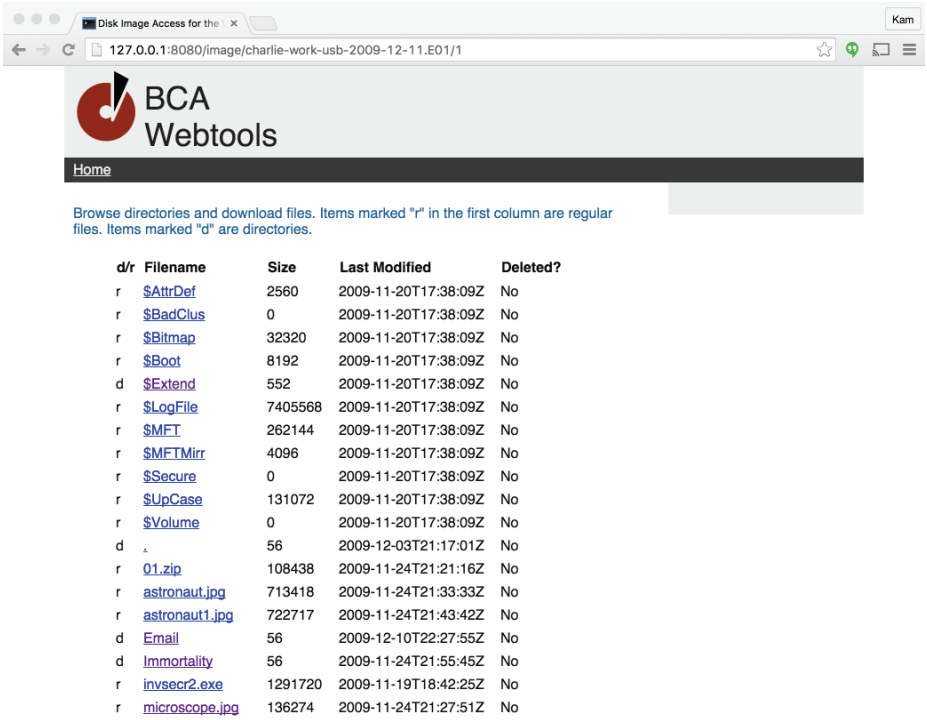


Figure 33 BitCurator presentation, Slide 44, BCA Webtools

Slide 45 (**Figure 34**): Gives instructions for how to download and install software from Github via BitCurator Access. It uses Vagrant to build a virtual machine and uses a local host.

The BitCurator environment is conceived of as a "working environment" and is intended as the lowest common denominator for archival workers. But all of the tools in BitCurator are also standalone open source apps that can be scripted and run independently of the BitCurator environment. It's their philosophy that users should not be confined to the environment if they develop the sophistication to do more with the tools and/or combine them with other tools not currently in BitCurator. A good example is the incorporation of some of these tools in the newest version of Archivmatica.

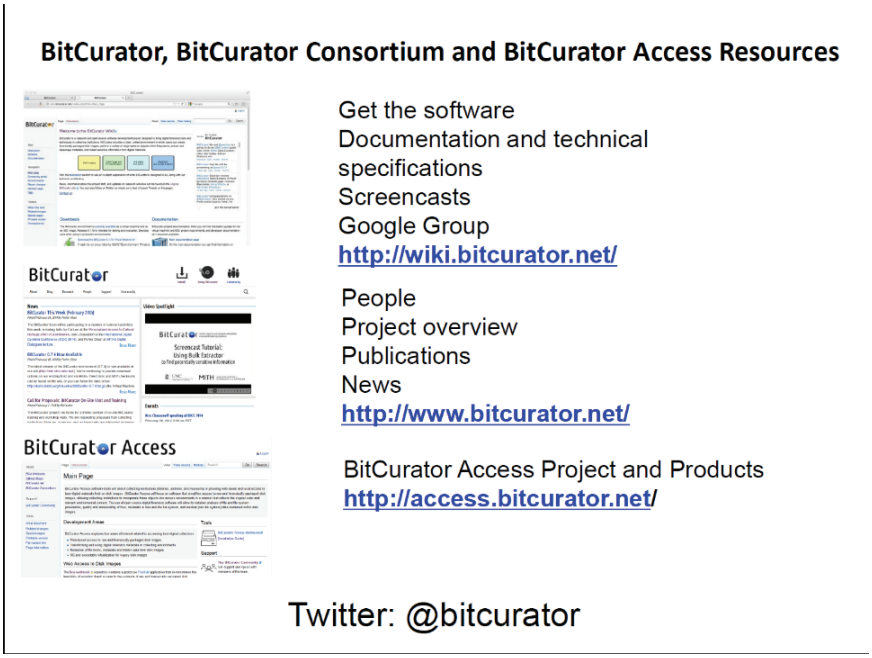


Figure 34 BitCurator presentation, Slide 45

QUESTIONS FOR CAL

Q*:* Do the tools have to be run against a disk image?

A: No, that's the default scenario, as most of these were developed for forensic work, but tools like Bulk Extractor can be run against directories or even individual files, or bags.

Q: How do you deal with a situation where the intellectual description of a collection spans multiple physical devices?

A: Cal's approach is that you simply grab what you need from the media objects or devices, which could be in high numbers, and sequentially just "plop" these files and/or images "somewhere" and then work on them later to make the decisions about groups.

JUSTIN SIMPSON — ARCHIVEMATICA

Justin discussed a recent project that Artefactual did for Simon Fraser University Archives. It was part of a larger project with 8 or 9 components, part of which included email. It centered on email accounts that the University IT dept. had taken off the email server and archived in the Zimbra format. The University staff had no knowledge of how to archive email, or tools to work with. Artefactual concentrated on how to get the email from the IT environment to the University Archives.

They had the IT dept. load the Zimbra archive onto a test account on their mail server. They then used an offline IMAP server to pull the email off the account into MailDir format on an Archivamatica machine. MailDir was used because that's how Archivematica does it -- it goes into the IMAP server and converts it to MailDir. They then converted the Maildir to MBOX -- one MBOX for every folder in the original email store, and then ingested the MBOXes into Archivematica.

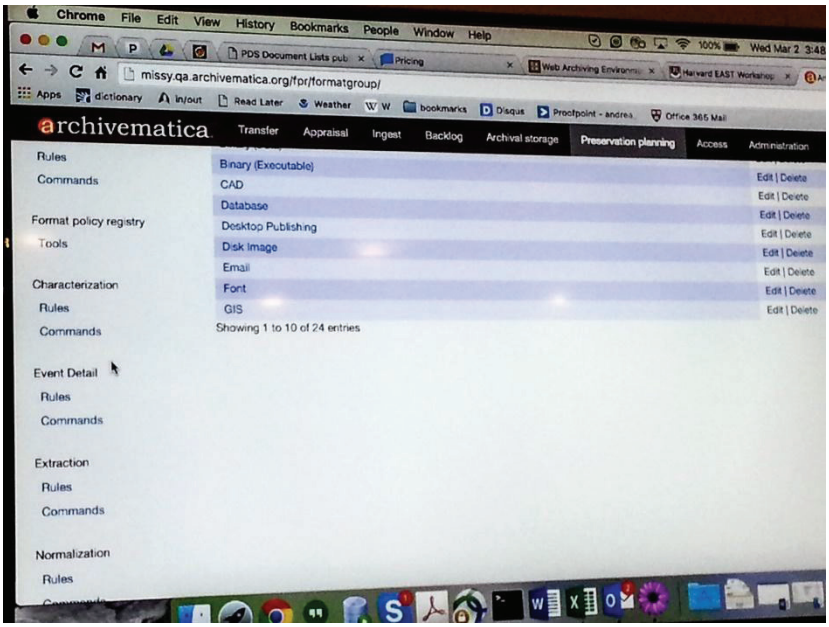


Figure 35 Archivematica Ingest Screen showing the MBOX import. The panel on the left shows the preservation tasks that Archivematica performs. These are based closely on the PREMIS event types.

Processing in Archivematica is based on

1. Identifying the format of the file. Archivematica uses a registry of format types to recognize the formats. These are very much based on the National Archives (UK) forms of MBOX. This is not a very detailed format description. Something that he would like to see discussed (which was not discussed at length in the succeeding period) is how to recognize different kinds of emails. At this point Archivematica id's file types by header ID and by extension, but they would like to have other methods to do this. Actions that are taken next are based on the identification of the format.

2. A typical next action would be extraction of the email files. The exact form of this action is dependent on the file type. Justin used the example of a 7Zip file. In the case of a disc image, they use components of BitCurator to extract the data. The purpose of this part of Archivematica is not just to provide a set of tools, but also to provide the commands for running these tools

Note: The identification of email formats was a subject of discussion on Thursday, in the context of the need for validation tools for sustainable email archiving formats

The commands are recorded as PREMIS agents for preservation events. So the metadata can record information such as "this command of this version of this tool was run by this person for this event"

Their work with the SF Archives focused on getting the material into a form in which it could be processed. They have been working to settle on the MBOX standard to reduce complexity by having a single format. MailDir was the format that Archivematica was originally created to work with in 2012, but by moving to MBOX, they will be able to make their system compatible with other systems, primarily ePadd.

They wrote a script that could go into an MBOX and extract all of the attachments. This Python script is linked as a command in the GUI, and now you can simply execute the attachment extraction as one of the built in commands. You can set up this extraction on an incoming MBOX file, which will allow you to run a variety of micro-services on the attachments: (**Figure 36**).

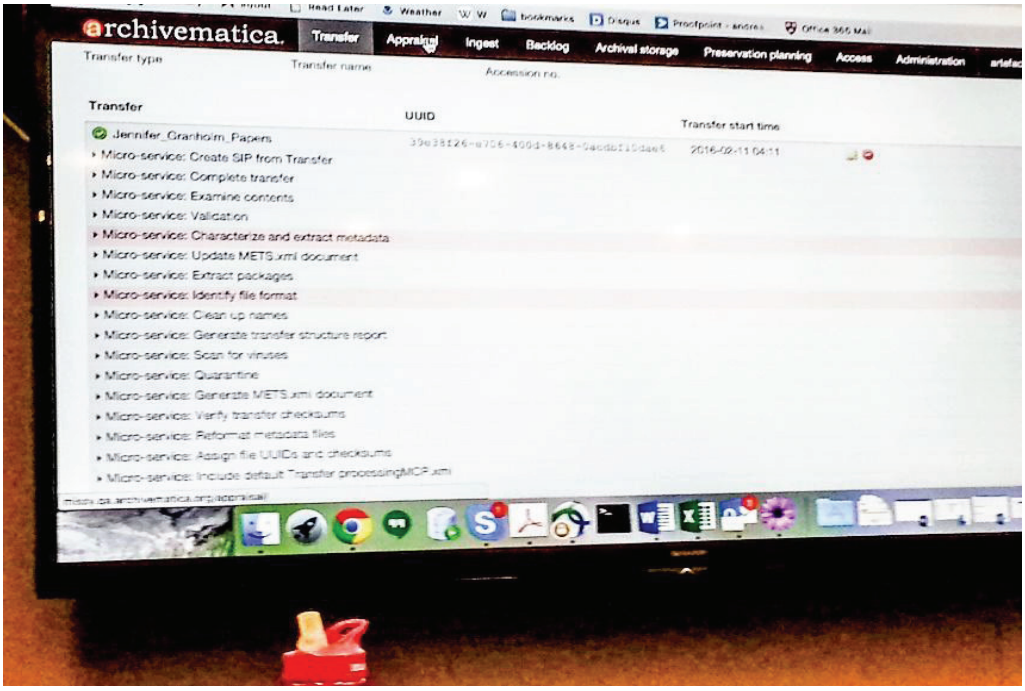


Figure 36 Archivematica Import screen, with micro-services for attachments

These are executed according to policies that are set up in Archivematica based on format. For example, you might have a rule that says that attached RAW image files should be converted to tiff using the appropriate micro-service, and then you convert to Jpeg for access.

Other things you can do: they have a micro-service called "examine contents" which runs Bulk Extractor from within Archivematica.

These features have been developed for the SF Archives, although all they have done so far is to characterize the files. Another example is a script that takes an MBOX file and generates statistics about it -- how many emails, attachments, who the authors are -- what they are calling technical metadata about the files. They don't have a good way to store that metadata.

Another area they are working on is better validation for the files. (Figure 37). They're using JHOVE. But better validation would allow them to distinguish various versions of a format in a precise enough way to invoke micro-services more efficiently.

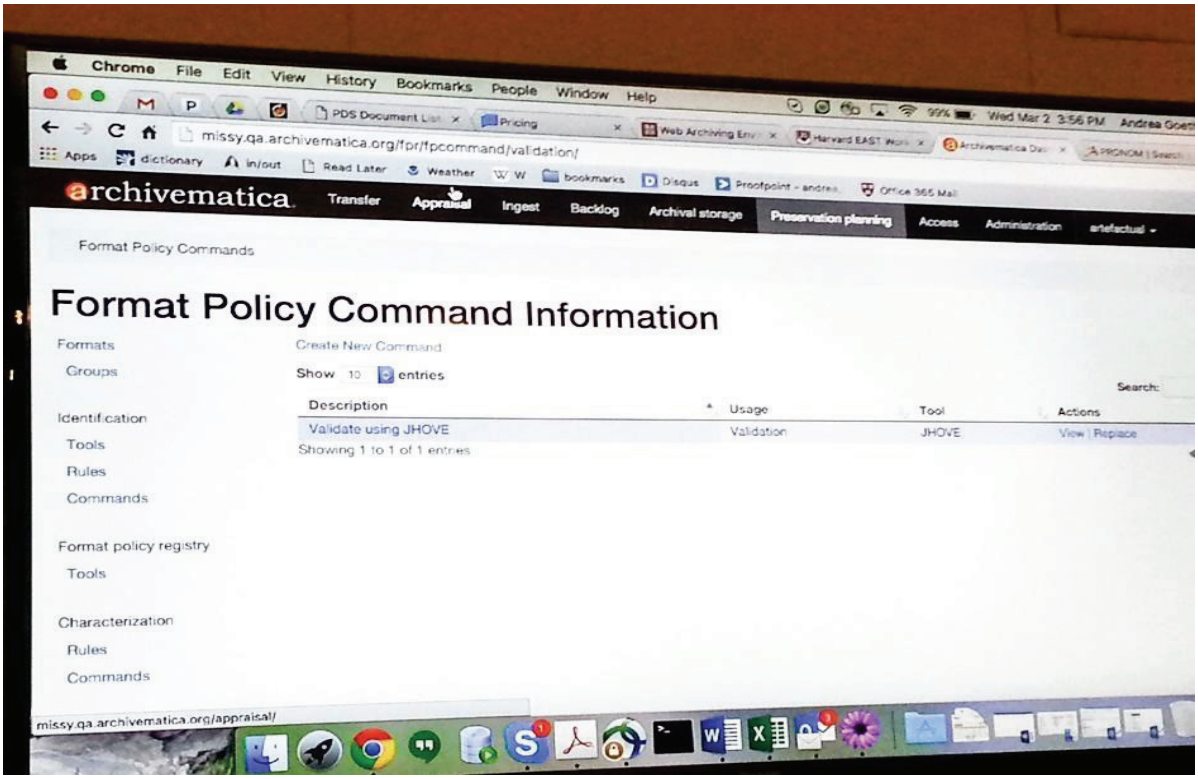


Figure 37 File validation in Archivematica

Appraisal functions. They got a commission to integrate Archivematica with DSpace and ArchiveSpace from the University of Michigan. The Appraisal tab in Archivematica is the result of this. The essence of this project is to connect the workflows of Archivematica and ArchiveSpace.

The first workflow step in Archivematica is TRANSFER, which consists of a set of micro-services that are read-only. It's designed to produce metadata. E.g., to generate new or verify existing checksums, assign unique identifiers, perform virus scanning, handle non-Unicode characters, and execute a number of the microservices that he had already mentioned. At the end of this process, you generate a SIP, at which point you make decisions about things like normalization, or generate derivatives. However, TRANSFER does not produce the SIP, but provides you with the data to make decisions about structuring the ingest, normalizing, and deciding how to handle different attachment formats, redact information, etc.

QUESTIONS FOR JUSTIN

Q: How do you show aggregate relations?

A: You can open an ArchiveSpace pane, which would or could have descriptions for the collection, and then drag transfers in Archivematica to the pane and drop them onto the description. The descriptions can be at any level. Each of these things will then become a SIP, which would be processed through the rest of Archivematica into an AIP. In the METS file for the AIP, all the relationships between the parts would be documented. The METS file's FileSec would contain the folder structure, and the relationships between the items.

The Archivematica AIP is a set of the METS files and the objects. The METS file has all of the preservation events recorded as PREMIS, a physical StructMap of the "physical" objects, and a logical StructMap, which defines the archival order.

Q: Is the idea that people would take the AIP they've created in Archivematica and transfer it to a long-term repository?

A: Yes. But Archivematica was originally designed to be a workstation solution. Users could produce AIPs even if they had no preservation repository to push them to. The AIPS were conceived of as "Time Capsules." This treated the AIP with its METS file as the preservation object, but METS is really intended to be a transfer record. The Michigan project will take the Archivematica data -- all of it -- and put it into a zip file, which will be attached to a DSpace item. It's an odd way of doing things, but necessary because of limitations in DSpace, which defines only 3 levels: Community, Collection, and item. You can't do description beyond the third level. So the solution is to put all of the original source data from the AIP into one zip file, and all of the administrative and tech metadata into another zip, and attach both to the item level DSPace record. Archivematica has a Storage Service which allows you to define where the AIP is going - DSpace, etc. These are developed as Plugins, e.g. Plugin for Fedora, DSpace, etc.

KATE MURRAY — LIBRARY OF CONGRESS

There are detailed talking notes in Kate's PowerPoint (http://bit.ly/22LwpcY)

Location of the "Chart:" (http://bit.ly/1RqXH11)

The most significant here are included for ease of reference:

Slide 2 (**Figure 38**): *The chart was originally created by Wendy Gogel. It presented a functional comparison of selected email archiving tools and services. It used color to great effect to convey a complex message. It contained just enough information – not too much, not too little. It was a chart one could share with non-technical people, including senior management, to help them understand what was done but more importantly, what still needed to be done.*

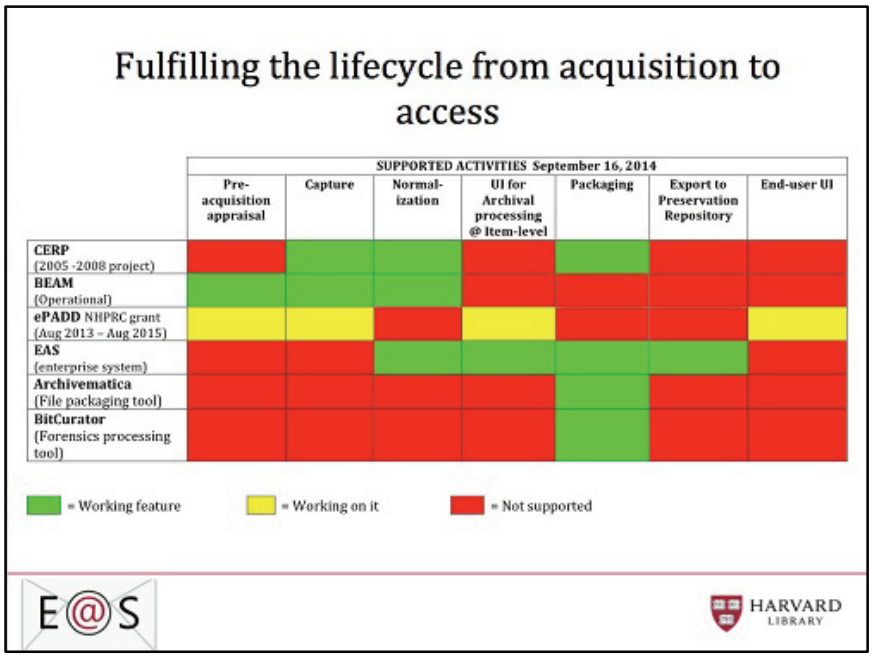


Figure 38 K. Murray presentation, Slide 2

It covered six common email archiving software tools, including of course Harvard's EAS and indicated their scope and maturity at specific lifecycle points. At a glance, one could understand the scope of a toolset at a high level and understand where a tool might need to expand to fill the need or other tools might be needed to pick up the slack.

Wendy shared the chart during the very first online meeting of the Email Interest Group in August 2014. Chris Prom subsequently used it in a blog post for the Signal in September 2014 (http://1.usa.gov/1REgtYd)

The ePADD group customized the chart to highlight the capabilities of their tool. Note some of the specific changes here not just in ePADD's capability – we see no more yellow for in development for them – but also the expanded lifecycle points, the start of some implementation factors and more tools are including in the comparison– especially commercial tools including Mailstore, AccessData FTK, ZL Unified Archive, and eMailchemy.

When the Archiving Email Symposium Workshop participants saw the chart again in June 2015, and the way that ePADD customized it, they loved the chart even more. Wendy said – let's work as a group to revise the chart to reflect new developments in tools and processing!

The chart was moved into Google drive and work began. http://bit.ly/1RqXH11. The group refined the lifecycle activities and added draft summary definitions – this is still very much a work in progress. More tools were added including DarcMAIL, Preservica, Paraben Email Examiner and Aid4Mail

Coverage was expanded to explain tool capabilities with both email messages and attachments

The option for "in development" was removed – it either does or it doesn't. And it would be impossible for us to know what was in development for commercial tools.

But we were moving away from the some of what made us love the chart to begin with. It's simplicity. There is the summary chart which we just looked at – nicknamed the color chart. We've come to realize that we don't always share a common understanding of terms so we've also started a glossary. The idea is that they'll be summarized on the color chart in the column headings but then have more in depth explanation in a separate Google drive sheet. This work is still in progress as mentioned. But we also thought we should address some of the other issues with the tools and services. So we made a companion page in Google drive to spec out the Cost and Implementation Factors.

Slide 13 (**Figure 39**): *The idea here is that more technically minded staff would be drawn to this type of information while the color chart would still serve the "overview" purpose.*

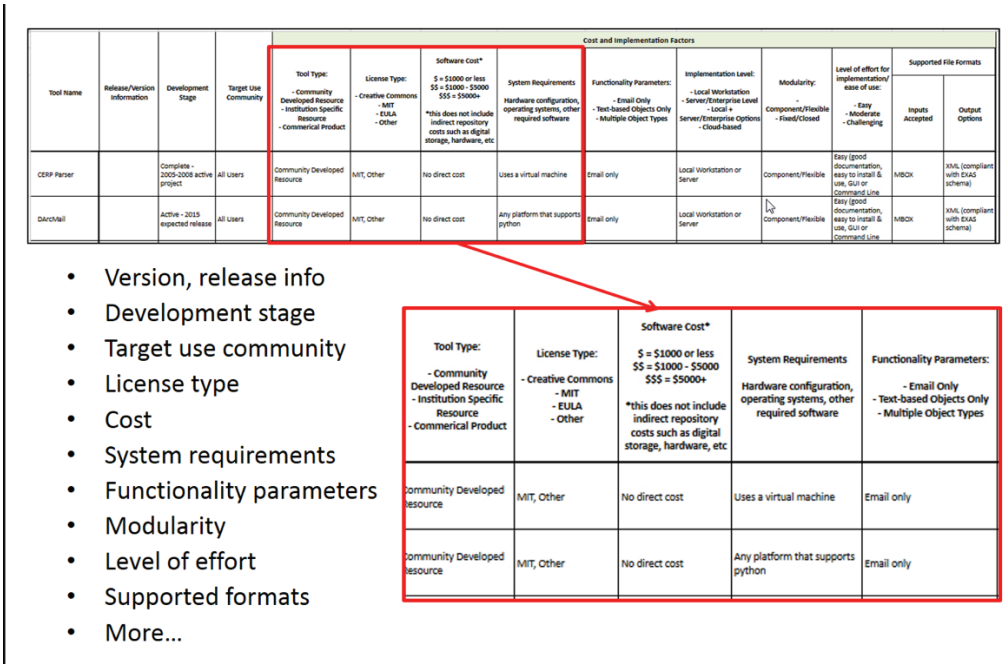


Figure 39 K. Murray presentation, Slide 13

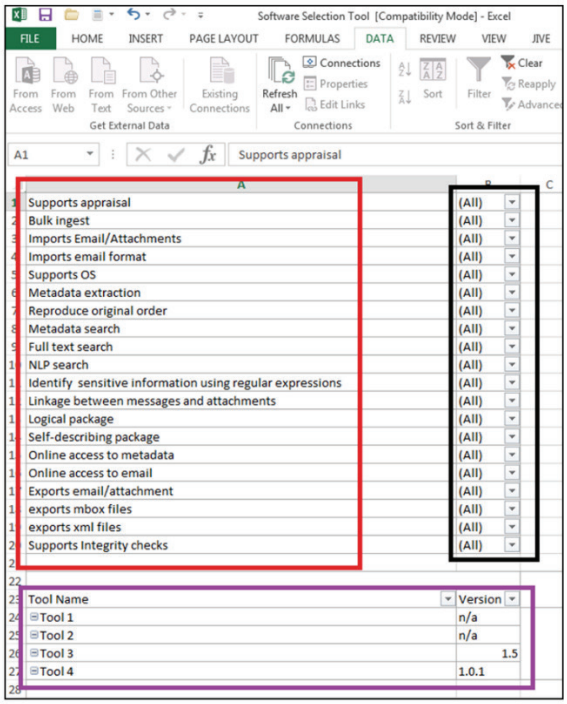
Word of the chart got out. The Email Archiving in a Curation Lifecycle Context panel at SAA 2015 used the chart to frame the discussion for a diverse range of topics from Glenn Edwards (ePADD), Ricc Ferrante (SIA), Wendy Gogel (Harvard Library) and Chris Prom (UIUC).

The color chart is perhaps most useful as a snapshot of the current overall landscape for archival email stewardship toolsets. At a glance, senior leaders and administrators could see where there are gaps with the tools currently in place within an institution so that resources could be allocated to further develop in house products or explore one of the listed software tools. Software developers and funding agencies could easily identify areas of opportunities for new toolsets or building on existing ones. Technical staff and digital preservation practitioners might find the cost and implementation factors useful in considering toolsets for existing environments. But some in our group thought there's more to the story. We need a chart that can help people decide what tools will work for their specific needs so…

Mark Conrad (NARA) developed the interactive Software Selection Aid for Archival Email Stewardship Software (which does not form an easy acronym SSAAESS) in late 2015. It's not really a chart at all but Excel workbook that relies on various filters and macros to help identify tools that meet specific needs. The Potential Tools sheet is used for gathering information about candidate tools that an institution might want to consider using. It's customizable so an institution could add whatever tools or criteria they'd like. The criteria in red are also listed on the Selection Aid sheet

*Slide 17 (***Figure 40***): The Selection Aid sheet allows the institution to analyze the data entered in the Potential Tools sheet by applying filters based on the criteria. At the top of column A (Red box) is the same list of criteria found in the column headings of the Potential Tools sheet. At the top of column B (Black box) are filters related to each of the criteria in column A. When this sheet is first opened all of the filters are set to (All). At the bottom of columns, A and B (Purple box) are the Tool Name and Version of tools from the Potential Tools sheet. When the sheet is first opened it lists all of the tools from the*

Potential Tools sheet.



Selection Aid for Archival Email Stewardship Software , Mark Conrad (NARA), 2015

Figure 40 K. Murray presentation, Slide 17

Slide 18 (**Figure 41**): *As filters are applied (Black box in Figure 3.) the list changes to reflect only those tools that meet the currently selected filter conditions (Compare Purple boxes).*

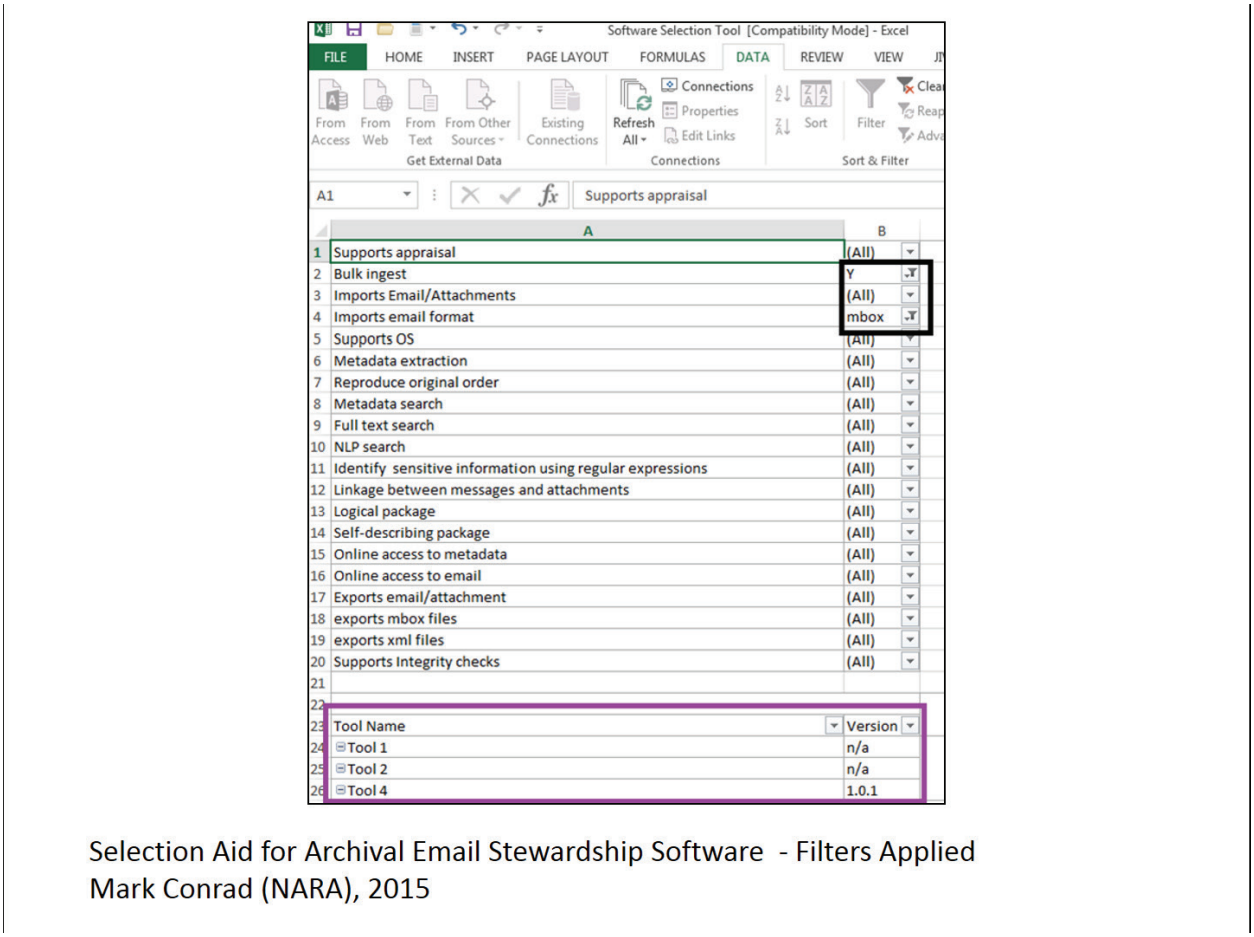


Figure 41 K. Murray presentation, Slide 18

Slide 19 (**Figure 42**): *So where are we? We're still working. When last we talked before the Christmas break, we agreed to regroup on how Mark's Chart would interact with the Color Chart – or are they really two different things with different goals and audiences. We also need to review our glossary and definitions and finalize our scope and context.*

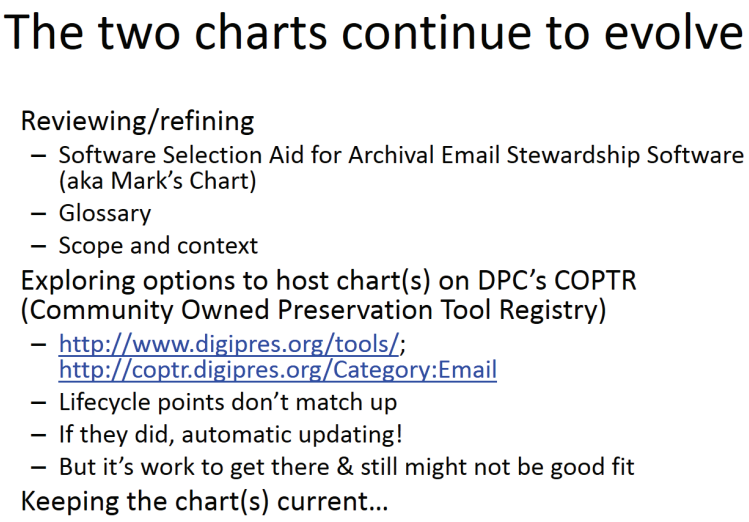


Figure 42 K. Murray presentation, Slide 19

Then what? Where does it live? Who minds it? Is there a permanent chart(s) wrangler? We're still working on that too. We'd like the community to have input and be able to customize for their own situation but how does that work, practically speaking. We've had conversations with COPTR about hosting it there but it's not at all a perfect fit. Mostly because their lifecycle points don't match up for email archiving. If we could map to theirs or even rework ours, the main benefit is that the chart would automatically update once relevant tools already listed in COPTR were updated. But part of the deal is that we'd have to agree to mind and develop the email tools in COPTR. So stay tuned. At some point, the chart(s) must forge out on their own but still be accessible and manipulable by the community. We've agreed to get the charts to a fixed point which demonstrates the current state of work but we, as an ad hoc group, can't maintain them ad infinitum.

QUESTIONS / DISCUSSION

Agreement to put Mark's chart on the Wiki.

In putting together the chart, has Kate developed a sense of what the best applications are overall, or does it depend on the context?

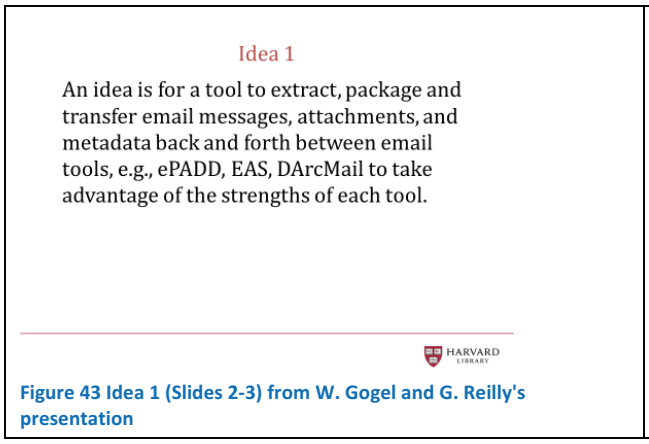
It does depend on the context. Certain applications will work better in different workflows. If you're trying to determine how a given tool will fit your specific needs, Mark's Excel spreadsheet is the more appropriate tool.

Ricc pointed out that the only problem with using Mark's sheet is that most archivists don't really know their own needs well enough to use the tool effectively, however, the exercise of trying to specify needs this way is very valuable for many people

WENDY GOGEL AND GRAINNE REILLY — HARVARD, USE CASES

Wendy discusses the potential of creating Use Cases to test ideas for different kinds of Workflow, and urges other members of the workshop to do likewise. She and Grainne introduce their Idea #1 as a way to kick off the discussion for the remainder of the afternoon.

The use cases are Illustrated in their Powerpoint: EAST-Workshop-Use-Cases.pptx, found at http://bit.ly/1UimtWW, although in the workshop they drew this on a whiteboard. In the PowerPoint, this Idea is illustrated in *Slides 2-3* **(Figure 43)**



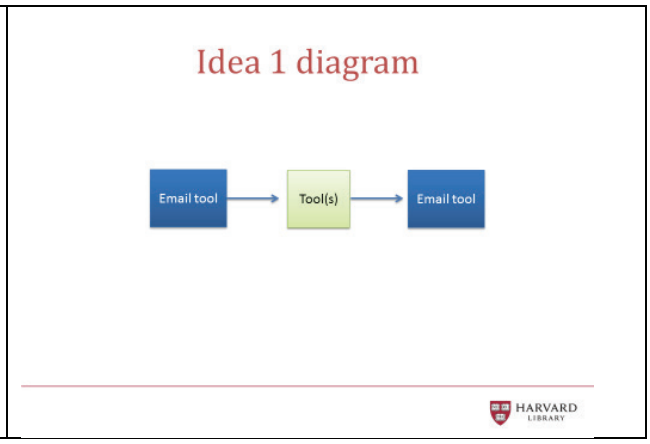
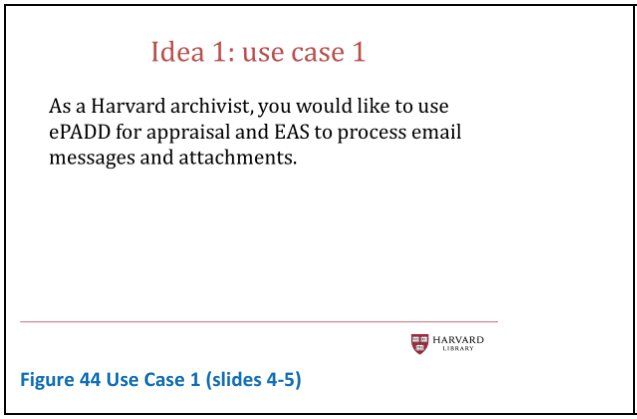


Figure 43 Idea 1 (Slides 2-3) from W. Gogel and G. Reilly's presentation

The first idea was for a tool to extract, package, and transfer email messages, attachments and metadata back and forth between email tools, for example, ePADD, EAS, DArcmail, Archivematica, and BitCurator.

A use case is presented as Use Case 1 on slide 4 and 5. The use case is for a Harvard Archivist using



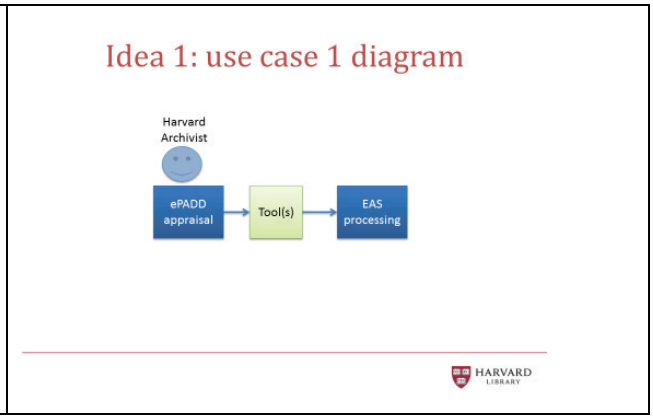


Figure 44 Use Case 1 (slides 4-5)

ePadd for appraisal and EAS for processing (Figure 44).

Following use case 1 for idea 1 the discussion became less formal. The whiteboard was used to cover various use cases rather than referencing the slides. The following day when the group returned to the slides it was noted that all the use cases in the slides had been covered during this discussion.

During discussions the use case expanded to include Use Case 2 as displayed on slide 7 (**Figure 45**)where a Smithsonian Archivist using DArcMail or a Harvard Archivist using EAS for processing would send the output to ePADD for delivery and/or discovery.

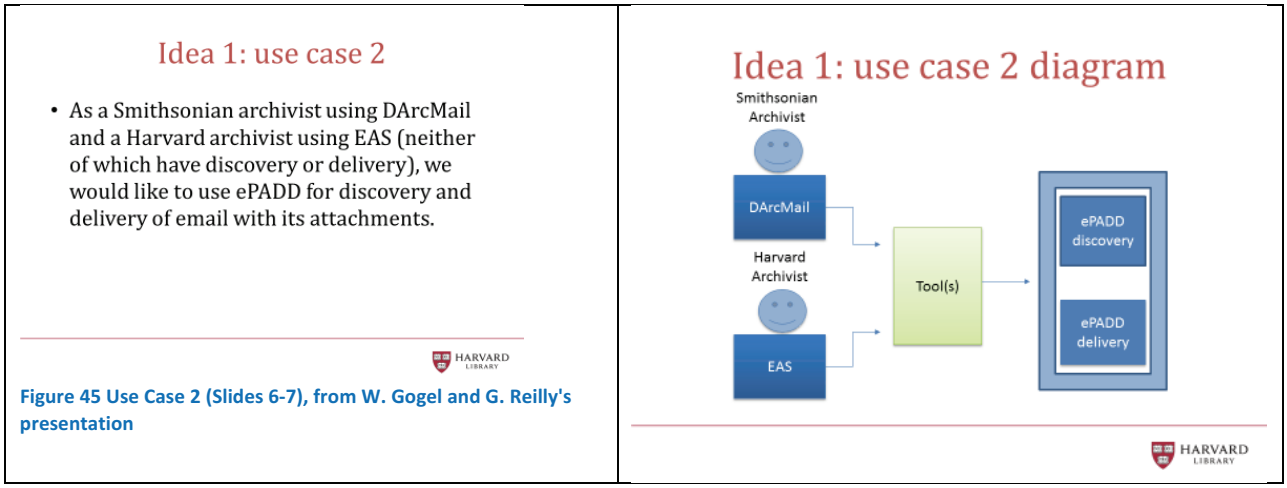


Figure 45 Use Case 2 (Slides 6-7), from W. Gogel and G. Reilly's presentation

The discussion further expanded to include Idea 2 as displayed on slide 9 (**Figure 46**). This use case includes sending output to a preservation repository, such as the DRS, from which it can then be delivered.

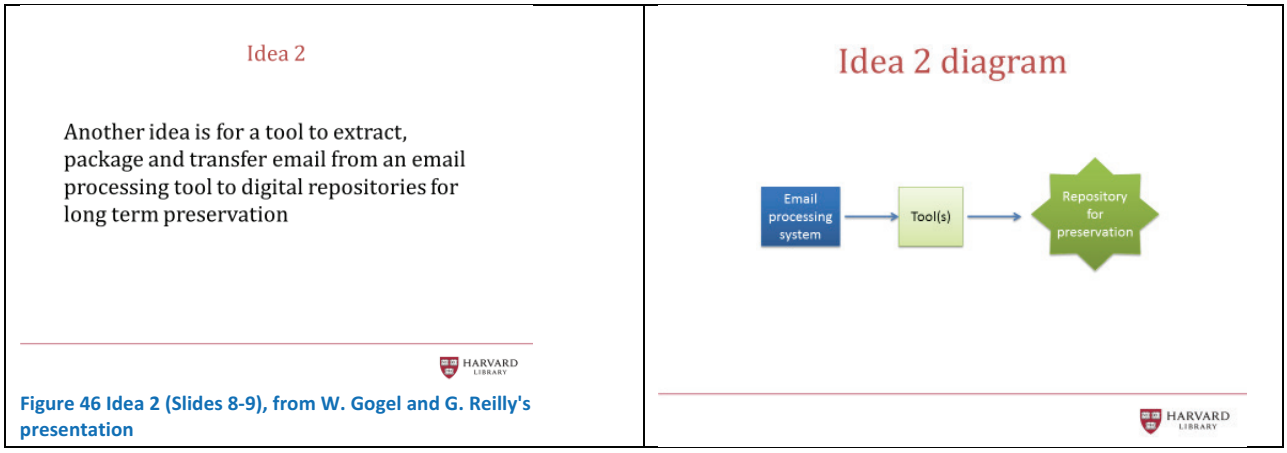


Figure 46 Idea 2 (Slides 8-9), from W. Gogel and G. Reilly's presentation

The discussion then generalized to cover Idea 2 use cases (slides 10 to 13) where output may be sent to the SI Archives Digital Repository or the Stanford Digital Repository.

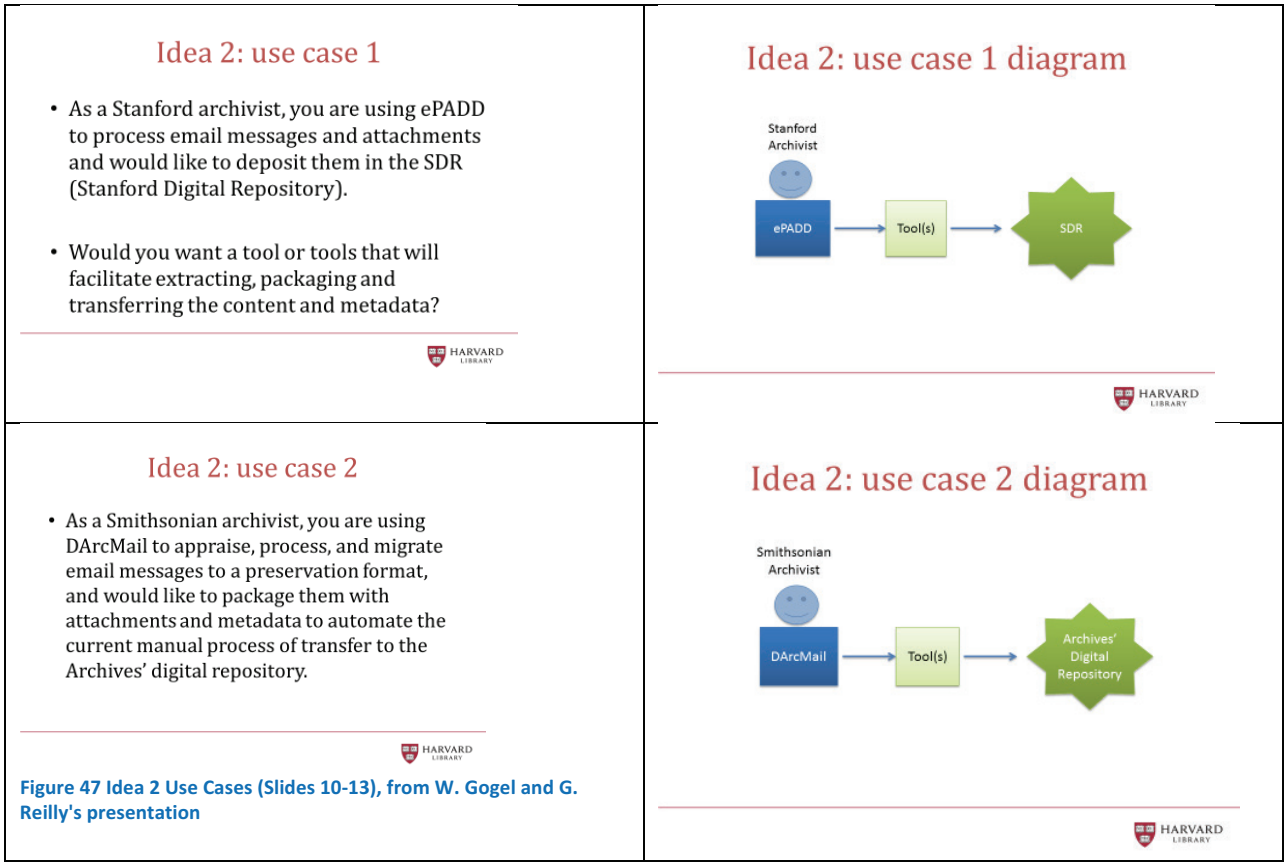


Figure 47 Idea 2 Use Cases (Slides 10-13), from W. Gogel and G. Reilly's presentation

Wendy sought clarification from Ricc about whether appraisal must be done before material is ingested into DArcMail.

Ricc stated that typically, this is true. DArcMail begins its process with acquisition, which would follow an appraisal process. He points out that ePadd has a much more powerful lexicon component for doing appraisal than what SI has available, so it would be desirable to use ePadd for the appraisal step.

If this were to be done, a tool would be needed to extract the data from ePadd and transform it for ingest to EAS (or DArcMail, in the case of SI). ePadd currently does everything internally, although there are plans to create an MBOX exporter for ePadd. How all the details of metadata management, and other structuring issues would be handled was noted by Glynn as "very good questions." The question then became whether it would make sense to have a tool sitting between ePadd and a target system, or whether it made more sense to incorporate export facilities into ePadd that could produce a somewhat generic output that was ingestible by other systems.

Kari suggested that once you could export from the ePadd appraisal system; you could use the SIP creation portion of Archivematica to format the packages into whatever form was needed for EAS or DArcmail.

The feasibility of this approach may hinge on clearer definitions of what constitutes the packages -- what metadata would be used, etc. Wendy suggested that the issue at this point is less about feasibility than whether the Use Case is valid. Is this a function that is needed, or would be found useful by many institutions?

Chris suggested that each of the tools represented in the use case have different strengths and are good at different tasks. For example:

- ePadd is very good at appraisal
- DArcMail is very good at preservation and the wrapping up of a digital object using an xml standard
- The EAS tool is good at metadata creation, and the pushing out of data
- Archivematica does a good job providing preservation services for attachments and the messages themselves

The question is whether people will agree to a single linear workflow in which you don't need to agree on the exchange standards, or whether there are so many different workflows, that people are going to want to go from one to the other. If everyone wants to go from one tool to another, you need to agree on a common exchange standard. To make the exchange between email processing tools work, you would need a very tightly controlled standard. Chris would love to see such a thing. The example that illustrates this is EAD, which is very difficult to move from one system to another because it's format is so flexible that more variation than can be easily accommodated in system designs is introduced.

Cal asked what this would look like. Is it just a superset of all the elements that might be interesting? If so, it could be gargantuan. Presumably the reason that you would prefer one of these tools over another is because it's adding elements that are not available in the others. It's not like DC, which boils the information down to a very broad common denominator. Ricc notes that it may not be as bad as that, as there is undoubtedly overlap between the systems.

Justin asked for clarification on the reasons for using ePadd for appraisal, with the understanding that EAS is needed to push data to the DRS? Is it because the email can be winnowed? Does this happen prior to import to ePadd, or is ePadd used to select (e.g) 50,000 out of 100,000 emails in an account? If so, if they delete email from the acquisition set, do they want a record of that deletion? Other participants agree that this is desirable. Glynn points out that a more typical scenario is one in which a collection of email is given to the institution, and then later the donor comes back and wishes to have some of them restricted or even struck from the set. In this case, ePadd will create metadata documenting these decisions. Accomplishing this kind of winnowing involves a combination of automated tools (as discussed earlier in the day -- the use of lexicons, and regular expressions) and manual labor to identify data that cannot be found using those methods.

Wendy asked Cal for clarification of his concern. Is the idea that they really want to move the rich metadata supported by one system to another, and is the problem that there may be cases where a

system doesn't recognize the metadata collected by another system. (*Note: Then, how does the receiving system manage data that it was not designed to manage?*) Cal's concern is defining the model of the data transfer between systems. Is it a data exchange, in the nature of DC where a common subset of data is carried forward, or a cumulative superset that potentially grows with each transfer between systems?

Chris notes that each system would not have to be fully aware of the data supplied by the transferring system, but there would need to be a minimal core set focused on an identifier for each message, so that actions for any particular message could be tracked, and some other minimal level of description for groups. The question is really what does each system need to know about the other in order to interoperate? It's not about the data *per se*. For example, one system might not need to know what lexicons were used to evaluate a message, but it would need to know that the message was flagged as restricted until further review.

Someone: *You could have two things where you had the core interchange metadata, and then an extended set of metadata. A system could take both, or just the core. The extended set might only be available as read only.*

Wendy asks if, in the instance that there is no overlap between two systems and a core dataset for interoperability was being used, would you have to lose that metadata, or would there be a way to store that metadata for use in yet another system, or another way of opening it up.

Ricc: I think if we define and document it well, people will have the information needed to make that call. The tools are sufficiently young that we may not know yet what we need. It may be that whatever a particular system delivers will be good enough for what is needed, and that not everything needs to be moved around.

Kari notes that if there are registries that list things like the lexicons used in different systems (and their versions, dates, etc.) then explicit information from an originating system might not need to be available in a receiving system. She basically suggests that detailed information about decision making policies and rules for processing should/could be retained in some centralized fashion for institutions (by each or centrally by all, not discussed) rather than duplicating this complex set of criteria throughout the repository. A system should simply be able to call on that.

Note: This re-introduces the problem Kari referred to earlier about the need, in such cases, to be connected to the web to use the tools.

Cal suggests that the problem is about conveying information from one system to another that is not preservation based, which is what PREMIS does. Right now, there is no formal way for conveying this "non-PREMIS" information. In the discussion there is a sense that the issues is about more than defining a schema for capturing this information, although that is a large part of it. The problem is to capture the context of a particular decision or action. E.g. on this date, in this institution, this policy was the reason that this particular decision was made.

Note: This could be handled by business rule, but there was some concern that this discussion could lead down a rabbit hole.

Determining the level of information, you need to carry forward from one system to another might depend heavily on your requirements for that information. There are some situations in which all you need is the content, e.g. just export an MBOX without any metadata. Then there are levels of needed metadata ranging from the very simple, such as an ePadd subject that's been automatically ID'd and mapped to an EAS tag. And then there are much more complicated situations which might require far more contextual information. Ideally you would have levels of metadata as options

Ricc suggests that what they want is something like a METS schema where they could plug in these various levels of metadata as needed. That kind of vehicle would give them the ability to have core metadata, but be able to pull in extensions as needed. Justin notes that what might be needed is a preservation planning ontology, that allows you to model the process by which a digital object is preserved, thus saving you from needing to carry forward a given version of the system that processed something, but rather allowed you to reference a canonical copy of that that system/policy/whatever in a controlled scheme. Right now, preservation policies are narrative, and therefore not easily amenable to machine processing. An ontology for describing institutional processing policies, procedures, and rules would make these easier to manage.

The first step may be to adopt a standard preservation format for the email itself, which at this point appears to be MBOX.

The PREMIS 3 standard does a good job of describing preservation actions, and this will account for many of the actions that need to be carried between systems.

There may be other types of information such as redactions that are not covered in PREMIS, but that probably should be, if not now, in the future.

There may be some level of descriptive metadata that doesn't work in PREMIS, but should be included in the metadata that goes with the objects.

The basic issues seem to resolve to the problem of standardizing the format of the email itself, which may be the easiest part, as MBOX is now the preferred standard, followed by the need to include preservation event documentation, which may be largely provided by PREMIS 3 or a subset agreed on as part of an exchange format, and then finally basic descriptive data elements that would be required and/or optional for exchange formats.

Moving this in a METS-like schema (perhaps even METS, although the participants did not wish to be prescriptive) would be the way to get data between different systems.

Another issue, assuming agreement on these approaches could be reached, would be to settle on a standard way of handling the MBOX and especially the PREMIS data — what syntax would be shared

among systems to ensure that data was not lost or misinterpreted? There are so many ways to implement these standards that agreement on one exchange method would be a major achievement.

A good example of the development of ontologies and schemas for managing data among systems can be found in the Hydra community. In this case, numerous interest groups are concentrating on defining the properties that they want to preserve for the content they manage.

Cal raised a concern about account level information that is not reflected in the MBOX standard. Any metadata standard would have to address this kind of object that is not immediately apparent in the structure of the content object. Ricc seemed to feel that some of these issues might be too esoteric, given the volume of material that he and other archives must deal with. Scale is a serious issue, and may preclude some of the finer -grained descriptive issues.

The definition of account was also discussed. It was deemed not to be simply the highest level of a folder or disk structure, but comprised all of the messages, links, attachments in-line and externally stored that comprise a logical account. In the case of DArcMail, an account is encapsulated in a single xml file which could grow to extraordinary size. Ricc points out that there are ways around this, but this is an aspect of scale that is a problem.

Much of the needed data is included in the MBOX file, so the problem may be how to determine what is not included (mostly account data?) and build or subtract out parts of a schema for that data, and include it along with the MBOX in exchanges between systems.

The good thing about this approach is that they're pulling pieces out of things that already work well.

The "tool" in the middle of the diagram is not necessarily a tool. It could be processes, policies, standards and practices. It's Magic! (Figure 48)

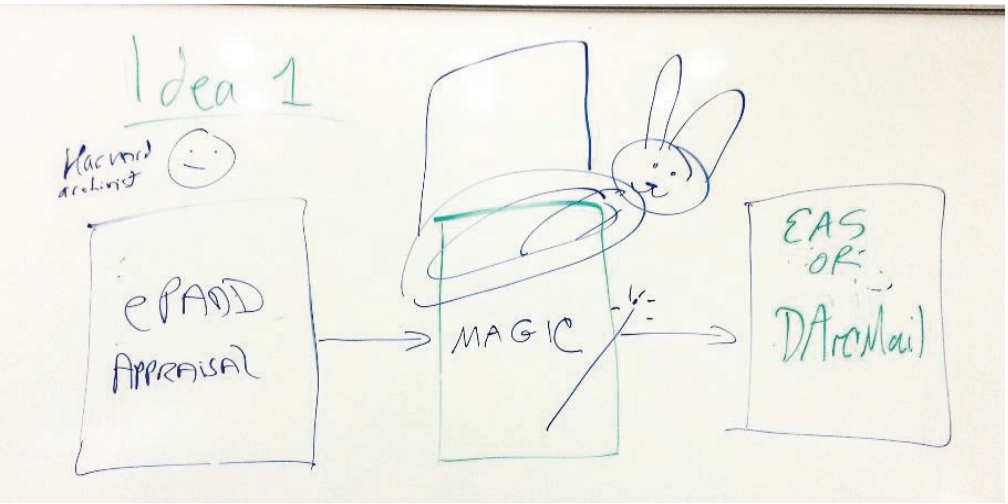


Figure 48 Discussion diagram from white board presentation

Someone: *"If they put some notes in there, when they were working in ePADD, and we can export that perhaps, into EAS, funnel those into a particular notes field, in non-public notes, say, in EAS, then that lets us continue to use their information, along with getting the email"*

Ricc, - and whether it's DArcMail or whatever, if we know how it's coming out, then we can build the receiving end of it.

Thursday, March 3, 2016

GROUP DISCUSSION

The morning began by discussing the topics that had been written by participants on the white board the previous day. Chuck asked that each person who wrote a topic explain it in further detail. Chuck proposed that after an initial discussion, the participants choose a few topics to concentrate on, and develop use case diagrams to more fully examine them.

Topics listed on the board were:

- Collaboration on lexicons:
    - for reuse - hosted by?
    - formatted - how?
- Collaboration/sharing of tool and scripts for transformation of email formats
- Use of Archival management systems as System of Record, how do email systems export to and from them?

Some additional thoughts that arose during the discussion were captured as “Parking Lot” Topics:

- Need for common terminology, e.g., for use in "the Chart"
    - "Transfer" vs "packet"
    - "migration" vs "normalization" etc.
- Curated vs. Records Schedule email

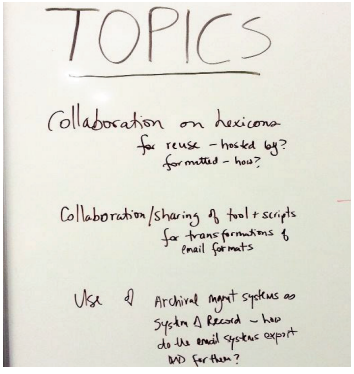


Figure 49 Proposed topics for Thursday discussion

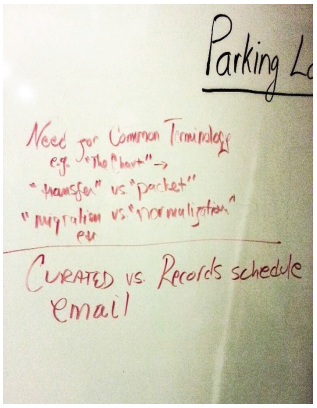


Figure 50 The "Parking Lot" for important, but deferred topics

Kari addressed the topics of Lexicons, tool and script sharing, and archival management systems as systems of record.

TOPIC: LEXICONS

Various systems use lexicons for search, for identifying sensitive information etc. ePadd and Bulk Extractor, for example both use lexicons. Can there/should there be collaboration in the creation of these lexicons? And if so, how would the resulting lexicons be managed? How do you make them available somewhere so other people or systems could use them? How would you format them so that they could be put in different systems?

TOPIC: SHARING OF TOOLS AND SCRIPTS.

The issue is the sharing of tools and scripts that people are using for the transformation of email. It relates to the Format Policy Registry that Justin showed yesterday (https://www.archivematica.org/en/docs/fpr) with the idea that "using this tool with this command produces this thing." Is there a place where we can share that kind of information? Where does it get housed? How would it be formatted?

Clarification (Wendy): *by transformation: conversion of one email format to another?* Kari: that was the original idea, but it can be broadened to include the idea of transforming and sharing tools and scripts.

TOPIC: USE OF ARCHIVAL SYSTEMS AND SYSTEMS OF RECORD

Archival systems are often used as systems of record, and the issue is how do the email tools and systems they have been talking about export metadata to and from these systems?

If, for example, an institution is using ArchiveSpace as their system of record, if they have email systems "out somewhere else," -- is there a main system into which all of these items should come, or would you have a segregated situation, where web archiving is in one place with all of its description and metadata, email is another place with its description and metadata? If you want to put at least some information into the general archiving system, what would be the kinds of exports out of the email systems that would be necessary to be pulled back into an archival system?

This idea is not restricted to description of the data. Wendy notes that if the issue is one of description, then ArchiveSpace could be another system added to the diagram in the "where it goes after the magic box"

An important point is that depending on workflow, different institutions may want to only move descriptive metadata, but not content from the email systems to an archival Collections Management System. The movement of metadata is probably easier than the movement of content, but content transfer can't be ruled out. Given this need to move descriptive metadata, the arrows in the diagram going from ArchiveSpace to the "magic box" should go both directions. (**Figure 51**).

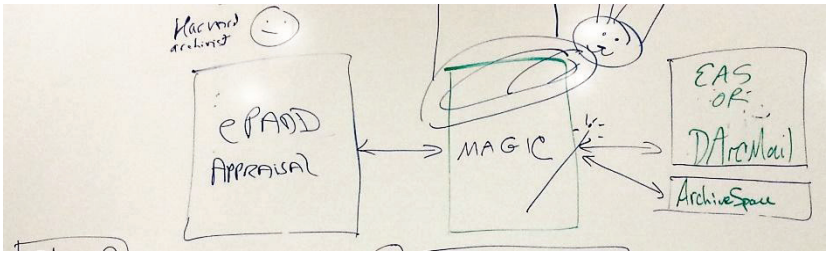


Figure 51 Modifications to the original discussion model

In fact, the bidirectional arrow implies that a Collections Management System might also be on the "left side" of the diagram, as an information source for an email tool.

The problem of sharing data between a Collections Management System and email processing tool, e.g. (ArchiveSpace and DArcMail) is that unlike sharing data between email management tools, this involves two different categories of system that aggregates information of different types and from different technological sources. The Collections Management System manages the full spectrum of archival holdings, and its description will be broader in scope. The lexicon issue comes into play when determining what metadata goes back and forth between the tools and the Collections Management System. Because the Collections Management System handles correspondence in the broader sense, you would want to be able to provide a description of the email portions of this, and then offer the ability to switch over to the email system to drill down more precisely in the dedicated email system. Preferably, depending on what the user needs to do with the content, and the type of email content itself, the Collections Management System in this situation would connect to different tools that performed the needed functions, e.g. instead of EAS, go to ePadd, or DArcMail.

The discussion seemed to imply that the "Magic" tool would create data for multiple outputs, particularly in regard to metadata (metadata more than content) so in some sense, it doesn't matter what systems are explicitly named as interacting with the "Magic" box. More specific use cases need to be developed to understand more precisely how this would work.

TOPIC: LINEAR VS. "MIX AND MATCH" WORKFLOW

The *Chart* seems to imply a linear workflow, and one of the initial ideas for the workshop was figuring out if there is a mutually agreed upon workflow for email processing and management. However, Chris points out that the discussion suggests that workflow varies significantly among institutions, and that there are typically important reasons why things are done differently in different places, therefore what's really wanted is a way to implement different tools in whatever order, or step in the workflow, where the functionality of those systems work best in a particular context.

It's important to understand the activities that people do with email preservation, but it may be that the order of these activities is different. An institution may move content immediately to a preservation platform, and carry out processing later. Other institutions may carry out a workflow more similar to the left-right arrangement of activities in the Chart.

Ricc points out that it will be important to develop concrete use cases that cover more examples of the processing order used by different institutions. So far we have looked at a Harvard hypothetical. Given the potential number of use cases, it's important to identify the ones that really need to be solved. Kari gives the example that she starts with Archivematica for the processing of the files, would like to pop that to ePadd for "processing."

In a number of examples that were brought up, the re-use of tools at different points in a workflow was discussed. A particular step in the workflow might be done with one tool, then content and metadata moved to another tool for further processing, and then back to the original tool for another step in the workflow. Although the order of functions performed might be the same, the tools used to perform these functions might have to be used in a different order, or used more than once as processing of items became more fine-grained or specific. This is a particularly critical issue for small institutions that may be looking for a setup that is close to turn-key, because they may have a less sophisticated sense of how the tools handle different activities.

Kari talked about her own workflow, in which there are currently no delivery and access systems, so no technically defined constraints on the structure of the AIPs. This allows her to concentrate first on file processing and then on content processing.

Kari diagrams this process on the white board (**Figure 52**)

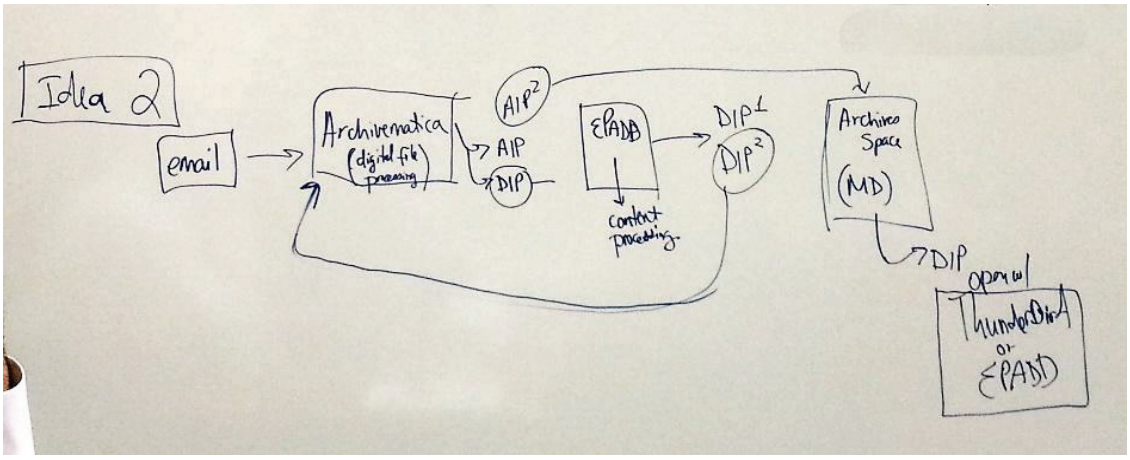


Figure 52 Kari's Diagram

An example, using the diagram: In ePadd, Kari creates DIP1, which is non-restricted information that can be made available immediately. DIP2 is the final version that has the trash and the "softball messages" removed. SIP2 could then go back through Archivematica to create an AIP, because the first time through the SIP consists of everything, but the second time through, it contains only the material she wants to keep in the form she wants to keep it in. Then AIP2 gets recorded back in ArchiveSpace.

Note: Local implementations can become complex with the introduction of archival collection management systems like ArchiveSpace for descriptive metadata.

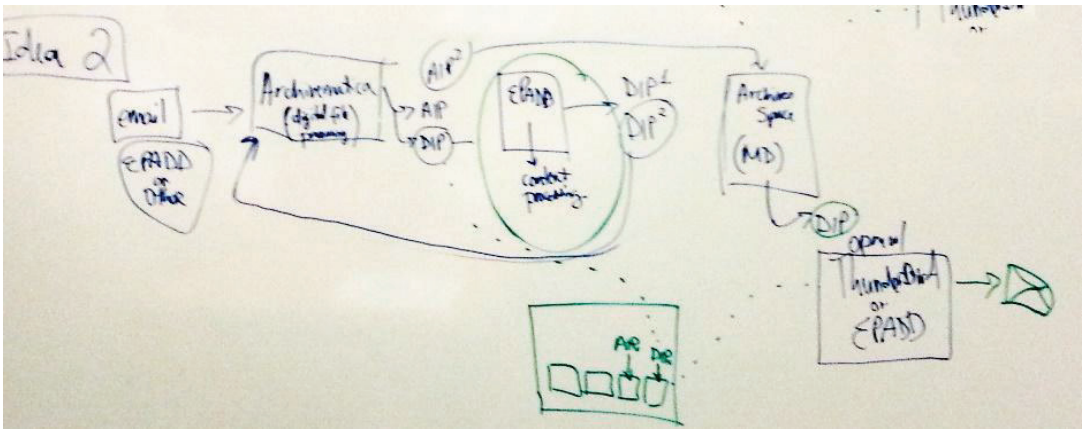


Figure 53 Additional modifications to model during discussion

1. Currently they do not have content processing for email
2. They get email in.
3. They run it through Archivematica.
4. Note that information in ArchiveSpace.
5. When someone needs it, she takes a copy, opens it in Thunderbird, and does reference out of that.
6. From Thunderbird she can export an individual message.

The problem is that it's too manual, and she can't do any of the content processing. She can discover and deliver, but she can't process the email. She discovers the aggregate of the email account through ArchiveSpace, and by opening up the email in Thunderbird and using Thunderbird's built-in search capabilities to find things.

The content does not live in ArchiveSpace, only the descriptions. The content comes from a DIP in Archivematica, and the description from a DIP from ArchiveSpace. The discovery that you can do in ArchiveSpace is at a collection or accession level. She can index emails, and manually type these into a notes field in ArchiveSpace, but this is both labor intensive and not good for searching or discovery.

Ideally, would they would want to start their high level description in ArchiveSpace, then move the high-level description back into Archivematica, and then have it all flow back into ArchiveSpace?

Not necessarily, because Archivematica provides more functions and services useful for analyzing email files, but it does not care about Descriptive metadata. In a situation where you have an existing collection to which you want to add an email component, you could start with the ArchiveSpace description and use it to seed the processing of email in Archivematica (or some other tool).

In the AIP from Archivematica, Kari can now say that a given set of digital content belongs to some digital collection, but just using an identifier of it, the accession number, for example. In this scenario, Archivematica would push only metadata out and not content.

Wendy reviewed the original use case idea -- to extract, package and transfer email messages, attachments, and metadata back and forth between tools. Although the original idea was to transfer all of these things, Kari's use case suggests that this doesn't need to be the case. You might want to transfer only certain items of metadata, and no content at all. And not even all of the metadata.

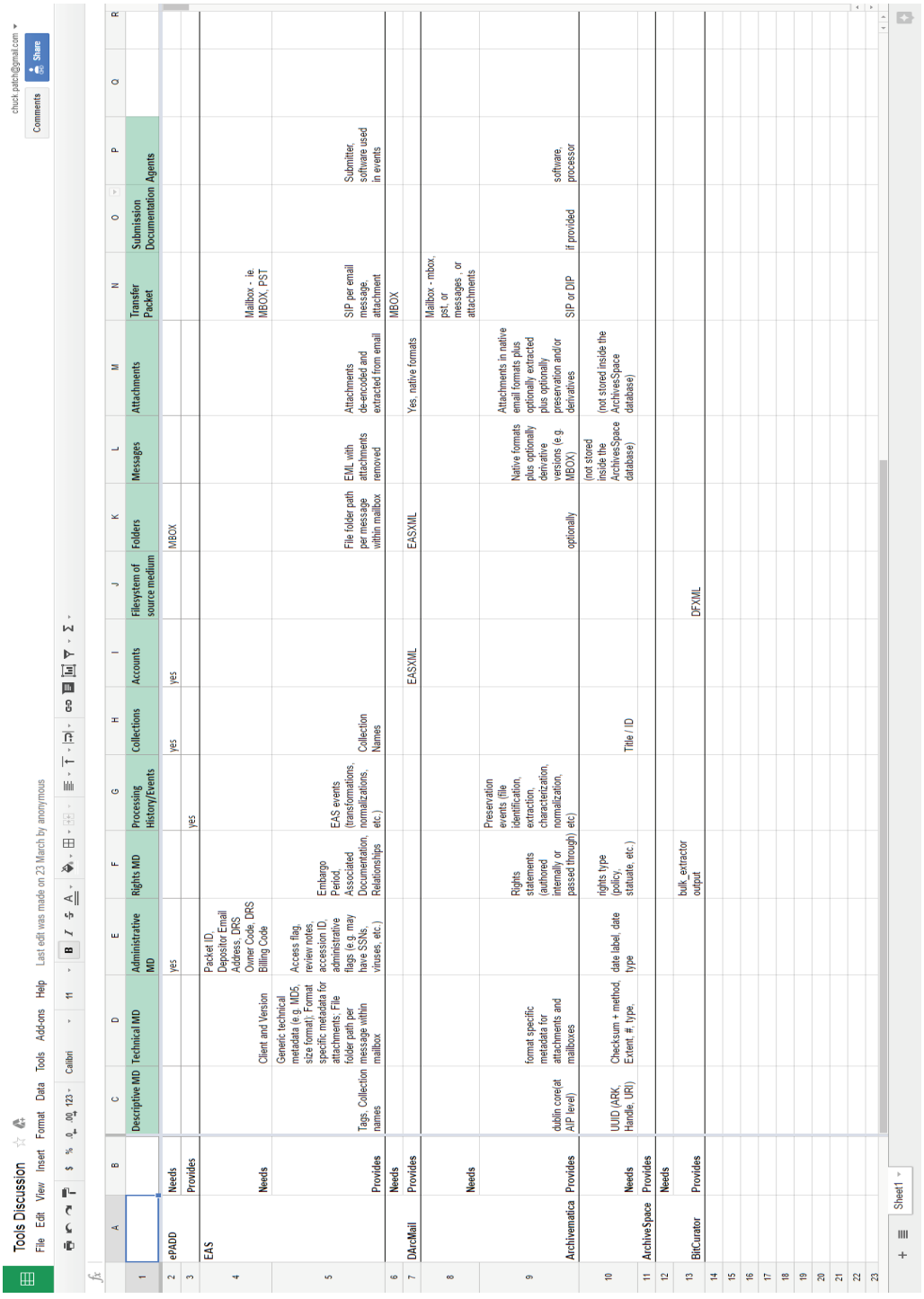
There was a discussion about where various tools would fit on the diagram, and which specific functions these provided. At one point, it was noted that no particular tool was used for just one thing, hence the circularity in some workflows between tools*, but not between functions*. Chris Prom suggested that this was becoming too tool-specific to be really useful, and that creating a table listing tools, with a column beside them listing the kinds of metadata they needed to operate would help clarify needs. This evolved over the course of discussion into a table described by Andrea Goethals and produced by Anthony Moulen. It is posted in Google Docs at https://docs.google.com/a/harvard.edu/spreadsheets/d/1v0JIeSzD_8GDXcceZMATCWJgrULu0VJrtuLp3Z4aHPw/edit?usp=sharing, and as shown in this document in **Figure 54** as it existed on April 22, 2016.https://docs.google.com/a/harvard.edu/spreadsheets/d/1v0JIeSzD_8GDXcceZMATCWJgrULu0VJrtuLp3Z4aHPw/edit?usp=sharing

This represented a narrowing in focus for the discussion. Instead of looking at the use of tools within a workflow, the approach first suggested by Chris and then picked up by other participants concentrated on the data required for email processing activities, and identified what various tools needed in order to work in each of these categories, and what data the tools were able to produce in these categories. This approach allowed the group to think of the data moving between systems as objects on which actions needed to be taken (although actions were intentionally not included in the table, representing a third variable that would be difficult to represent.) It was pointed out that this moved the focus to defining the boundaries of the tool systems by showing what went in and out of them, thus making it easier to figure out how they could be moved around in a workflow.

Additionally, the table seemed to help look at the flows of various objects as separate entities, which is often important in email processing, where metadata, attachments, and messages might have different flows depending on the required processing, and then be stored and accessed in different places, or brought back together. Moreover, they don't necessarily move through a workflow at the same rate -- certain parts, or types of material in a collection may be processed more quickly than others. In other words, even within the variability of workflows across institutions, within any given institution, the workflow is not necessarily monolithic.

Fuller definitions of each of the column categories should be developed, and perhaps simply included as annotations for the headers. For example, there was substantial discussion of the meaning of "Submission Documentation," which was ultimately defined as contextual information surrounding the digital objects and how they arrived at the repository, but NOT necessarily data about the content itself. Not really metadata, but lots of different kinds of data. The notion of creating a glossary for the terms used in the table came up.

In terms of designing the structure of an exchange, one of the problems recognized with the table is that some of the objects are metadata, and some are not. An effort was made to keep the concepts at a fairly abstract level, although there was difficulty in nailing down some of these concepts.



		Descriptive MD	Technical MD	Administrative MD	Rights MD	Processing History/Events	Collections	Accounts	Filesystem of source medium	Folders	Messages	Attachments	Transfer Packet	Submission Documentation	Agents
ePADD	Needs			yes			yes	yes		MBOX					
	Provides					yes									
EAS															
	Needs		Client and Version	Packet ID, Depositor Email Address, DRS Owner Code, DRS Billing Code									Mailbox - ie. MBOX, PST		
	Provides	Tags, Collection names	Generic technical metadata (e.g. MD5, size format), Format specific metadata for attachments, File folder path per message within mailbox	Access flag, review notes, accession ID, administrative flags (e.g. may have SSNs, viruses, etc.)	Embargo Period, Associated Documentation, Relationships	EAS events (transformations, normalizations, etc.)	Collection Names			File folder path per message within mailbox	EML with attachments removed	Attachments de-encoded and extracted from email	SIP per email message, attachment		Submitter, software used in events
DArcMail	Needs							EASXML		EASXML		Yes, native formats	MBOX		
	Provides												Mailbox - mbox, pst, or messages, or attachments		
	Needs														
Archivematica	Provides	dublin core(at AIP level)	format specific metadata for attachments and mailboxes		Rights statements (authored internally or passed through)	Preservation events (file identification, extraction, characterization, normalization, etc)				optionally	Native formats plus optionally derivative versions (e.g. MBOX)	Attachments in native email formats plus optionally extracted plus optionally preservation and/or derivatives	SIP or DIP	if provided	software, processor
	Needs										(not stored inside the ArchivesSpace database)	(not stored inside the ArchivesSpace database)			
ArchiveSpace	Provides	UUID (ARK, Handle, URI)	Checksum + method, Extent, #, type,	date label, date type	rights type (policy, statuate, etc.)		Title / ID								
	Needs														
BitCurator	Provides				bulk_extractor output				DFXML						

Figure 54 Table developed during the workshop, as of 4/22/2016

Note: Wendy and Grainne are trying to clarify the organization of the table in a second version to be distributed after the Workshop.

One example of confusion is the term “collection,” which is interpreted differently by people -- it seems to convey a combination of content and metadata, but the proportionality of those components can vary. In general, it became clear that columns toward the right side of the table refer more to content, and the left side more to metadata. Where the dividing line between content and metadata lies seems to vary by institution and workflow. Some places, eg. Harvard, regard email messages and attachments as the sole content. For other institutions it could include folders and accounts. For example, the content of a .pst file includes folder and account information -- these structures do not have to be generated as metadata, although this might be done preferentially to provide greater power in filtering, searching, or combining the mailbox with other content sources. Problematic issues include things like orphan attachments.

The Needs and Provides rows for each system were described as follows:

- "Needs:" this is the minimal set of data that is required to transfer something into the system.
- "Provides:" is what a system can output -- the maximum that the system can output.

Although the decision was made not to include the row in this iteration of the table, a row for "could" might lay the basis for development projects.

The EAS row was entered in more detail than the others as an example of what the table should look like when fully populated. In this process, it became clear that the needs are often driven by policy and administrative necessity, as well as technical constraints.

It was acknowledged during the creation of the table that many of the column definitions were fuzzy and would need clearer criteria. Another issue: depending on how a workflow is created, you may be in a situation where a particular system does not "need" a certain piece of data, but a system later in the workflow does need that data, and thus it has to be passed on. Is the data that will be passed on listed in the *Provides* row, or should it stay in the *Needs* row? How can one account for a situation in which the system loses data that is passed from a previous system?

There was much discussion of nuances -- e.g. the reconstruction of folder structures in Archivematica can be done by inferring it from the pathnames, but that means that the location of the folder information is in the messages. The majority of the discussion appeared to apply to specific characteristics of the tools and how they were used in specific circumstances.

Note: The continued population of the table, and the creation of supporting materials -- glossary, and criteria for filling in the cells -- should be continued when Wendy and Grainne distribute the new version.

TOPIC: VALIDATION TOOLS FOR MBOX AND EML

This is a more specific issue, but addresses a need because of variability in these standards. It's probably easier to build for EML, because it's based on RFC5322. MBOX is less structured, so more challenging, but someone could put feelers out to the community about how that could be built. How would it be implemented? Would it be a component of JHOVE?

This is something that would be particularly useful. Yesterday there had been general agreement that the MBOX format is becoming the standard, but is the MBOX produced by ePadd something that could be easily imported into Archivematica? Is it lossless? This is a topic that would also be of interest to the greater community. They would stand to benefit even if they are not using the tools discussed at the workshop. If there were a way to validate a preferred implementation of this format, it would allow the greater community to go to vendors who produce email management tools and get them to respond to the needs of this greater community.

Given the agreement that this is an issue for the greater community, Kate volunteered to take this up with LC. LC is exploring funding issues, perhaps in collaboration with NARA.

Kari brought up the more basic issue of format identification, and what could be done to develop systems that could identify formats by signatures and not just rely on file suffixes or header ID's. Kate points out that there are multiple sources of file format information that don't necessarily map to each other, such as the LC Sustainability website and NARS's Pronom. In fact, the library doesn't even use its own file identification work, preferring instead Droid and Jhove.

Someone pointed out that in Pronom, the signature for MBOX is a file extension, even though the header contains identifying information specific to the MBOX format. Even being able to identify the genus MBOX, and not all 4 variants of it would be helpful, although validation utilities that could do that would be more helpful.

Wendy presents a second use case involving a Stanford archivist (**Figure 55**). It seems to fall into the same generalized use case we have been talking about.

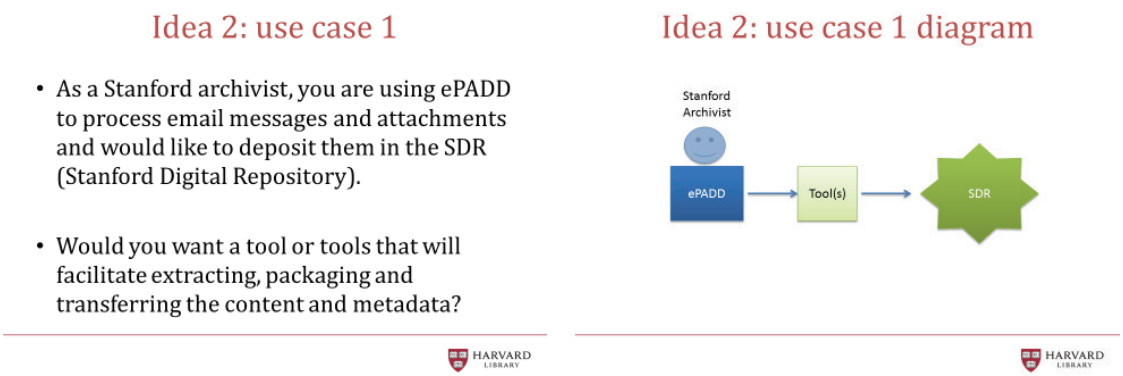


Figure 55 Stanford Use Case

Ricc asks if those also working on non-email digital objects already have an automated solution for *that*? Most places do not, and there is discussion as to the extensibility of tools that are developed for email. What are the common issues among different types of digital object that might be important in the further development of tools for email and the future development of archiving tools for things like web sites? A key issue is how much information these systems need to carry forward -- is it enough just to bag everything in a manifest and export it, or does there need to be more information about the content? Several participants said that for long-term preservation this was vital, but there was some disagreement about the extent to which any given system needed to be aware of, or be able to use in a meaningful way, the data it received.

Ricc reframed this question in terms of the repository, which he described as a closet. He needed to put a bunch of stuff into the closet, making sure that their relationships stayed intact, but the closet didn't need to be aware of the individual items in the set of things. This led to a discussion of how AIPS were produced and used. The issue revolved around how much the repository needs to "know" to support the amount of processing desired for long term preservation of objects in the repository. Harvard supports the idea of knowing a lot for their repository so that they can take preservation actions within the repository (sometimes global actions). In a more practical sense, could modifications be made to the AIP directly using external tools, or was it necessary to remove the AIP as a DIP, make the modifications and re-ingest the AIP? Ricc described his work as burying the AIP in the repository and "exhuming it when it needed to be changed. Glynn and Kari also described taking the AIP out, working on it, and putting it back into the repository.

DISCUSSION SUMMARY

Chuck gave a 5-minute summary of the workshop outcomes:

The most important outcomes for the meeting were that the participants agreed on 3 areas that are really important for the community to continue working on:

1. Continue exploring the idea of creating tools to extract, package and exchange content between systems.
    a. Next steps for this work are for everyone to add to, expand and update the Tools Discussion spreadsheet once a new version is distributed by Wendy and Grainne
2. Explore methods of sharing lexicons that can be used by the multiple tools
    a. Methods for accomplishing lexicon sharing are being worked on by an ePADD group, which includes Kari Smith.
3. Develop tools for identification and validation of sustainable formats for email.
    a. Next steps are for Kate to take this idea back to LC

To wrap-up this portion of the workshop, Wendy declares victory and provides the group with information about the closing session, an open-meeting for all of Harvard Library and EAST Workshop participants at the Forum Room in Lamont Library:

EMAIL ARCHIVING IN A CURATION LIFECYCLE CONTEXT: A PANEL PRESENTATION

From 1:00 – 3:00 on Thursday, sixty-five people attended the panel discussion entitled "Email Archiving in a Curation Lifecycle Context," which closed out the workshop. Christopher Prom served as moderator and panel speakers included Glynn Edwards; Riccardo Ferrante; and Wendy Gogel. A WebEx link (https://youtu.be/EFhRP7rJIhMM) to the presentation is available.

Brief presentations covered components of the curation lifecycle for email, stewardship functions (pre-acquisition appraisal, accessioning, processing, access, and preservation), issues, gaps, and tools. Originally presented at the Society of American Archivists in August 2015, the panelists updated their content to reflect developments over the past six months.

CONCLUSIONS

The community is very interested in working together to solve the problems we face. We agree on some of the essential needs, which helps us to set a direction for future work. These include the need to share controlled vocabularies used by the various tools; the need for identification and validation of sustainable email formats; and the need for an exchange standard that enables interoperable ways to extract, package and transfer data between tools.

There was an additional turning point for the community during the workshop where we recognized that we had conceived of a singular, linear (and potentially monolithic) workflow that we hoped to identify and agree upon. However, we realize that a sustainable approach may need to be more flexible. There is no standard workflow for everyone and therefore the same tools could be combined and used differently. The new concept recognizes that diverse institutional requirements and workflows need to be met and therefore, we are more likely to meet those needs by concentrating on the strengths of individual tools and taking a modular, or mix-and-match, approach.

SUBSEQUENT COVERAGE

Information about the workshop and its results were disseminated subsequently:

EMAIL ARCHIVING STEWARDSHIP WORKSHOP ON THE HARVARD LIBRARY BLOG

- On March 16, 2016, Harvard Library published an article written by Harvard Library Communications in the Library blog: *Email Archiving Stewardship Workshop* at http://library.harvard.edu/03092016-1642/email-archiving-stewardship-workshop. (*See pp. 71*)

NATIONAL DIGITAL STEWARDSHIP ALLIANCE (NDSA) STANDARDS AND PRACTICES WORKING GROUP PHONE CALL

- On March 21, 2016, Kate Murray and Wendy Gogel contributed a summary of the Harvard EAST Workshop to a discussion about email archiving as part of the National Digital Stewardship Alliance (NDSA) Standards and Practices Working Group phone call (http://ndsa.org/working-groups/standards-and-practices) with Mellon Foundation representatives.

O EMAIL! MY EMAIL! OUR FEARFUL TRIP IS JUST BEGINNING: FURTHER COLLABORATION WITH ARCHIVING EMAIL ON THE SIGNAL – THE LIBRARY OF CONGRESS DIGITAL PRESERVATION BLOG

- On May 10, 2016, Kate Murray posted on The Signal — the Library of Congress digital preservation blog: *O Email! My Email! Our Fearful Trip is Just Beginning: Further Collaboration with Archiving Email* at http://blogs.loc.gov/digitalpreservation/2016/05/o-email-my-email-our-fearful-trip-is-just-beginning-further-collaborations-with-archiving-email. (*See pp. 73*)

Email Archiving Stewardship Workshop

In this one-and-a-half-day workshop, email archivists from across the country gathered at Harvard Library to share tools and strategies.

See Also:

- **Harvard Library Blog**

From left to right: Franziska Frey, Christopher Prom, Glynn Edwards, Riccardo Ferrante, and Wendy Gogel. Photo courtesy of Kari Smith.

On March 2 and 3, practitioners of email archiving from multiple cultural heritage institutions such as the Smithsonian Institution Archives, the Library of Congress, Stanford University Libraries, and MIT gathered together at Harvard Library for a workshop on email archiving stewardship tools. “Born-digital archiving of any kind involves a lot of technology and skill,” said Wendy Gogel, manager of digital content and projects and leader of the workshop.

The goals of the workshop were to foster the expanding email archiving community, share updates on current work, identify needs for upcoming work and future directions, and expose the Harvard Library community to the issues involved in email archiving.

Participants found that collaboration is key to tackling the challenges in this field. “Rarely can one institution take on independently all of the work for every format,” Gogel said. By working together on interoperable open-source software, institutions can learn from one another and build workflows around the strengths of the tools being developed—regardless of who developed them. As more people use and contribute to the software over time, the maintenance of the tools becomes sustainable.

The community of email archivists is dedicated to working together to solve problems. Stakeholders agree on essential needs that help set a direction for future work, such as sharing the controlled vocabulary used by various tools, the need for tools to validate sustainable email formats, and the need to develop ways for email archivists to extract, package, and transfer data between tools. Attendees were also able to view tool demos and hear updates on work from Stanford University, the Smithsonian Institution Archives, Harvard Library, the University of North Carolina at Chapel Hill, the Library of Congress, the University of Illinois at Urbana-Champaign, and Artefactual Systems in Vancouver.

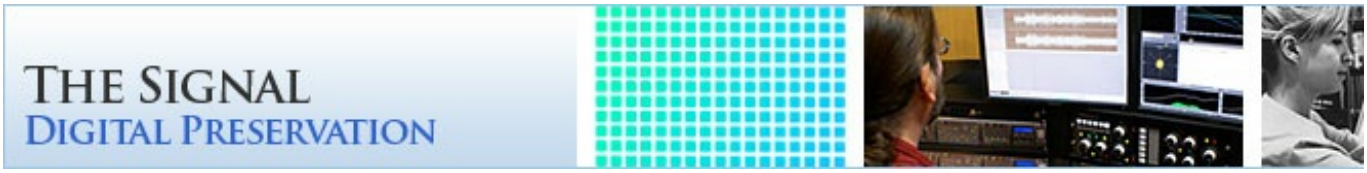
Scale and privacy are two of the major challenges for email archivists. Email accounts, whether collected as institutional records or contributed by donors, can dwarf the size of other collections—especially when you include attachments. Privacy is also a concern. Having one’s emails considered for scholarly research is a thorny subject; it may be years before some emails can be released to the public, and even work email may contain sensitive personal information or subject matter in unexpected places—for example, Social Security numbers and credit card information. Even with permission to release emails from a primary account holder, email contains third-party information by the other correspondents. "Email is one of the richest, one of the most revealing, if not *the* most revealing, of sources currently being generated,” said Christopher Prom, assistant university archivist and associate professor of library administration at the University of Illinois at Urbana-Champaign.

Franziska Frey, Associate Librarian for Preservation and Digital Imaging, welcomed the group on the first day and opened the public panel presentation on the second day. Sixty-five people attended the panel discussion entitled “Email Archiving in a Curation Lifecycle Context,” which closed out the workshop. Christopher Prom served as moderator and panel speakers included Glynn Edwards, head of the technical services division in the

Department of Special Collections and University Archives, Stanford University; Riccardo Ferrante, information technology archivist and digital services program director, Smithsonian Institution Archives; and Wendy Gogel. A WebEx link to the presentation is available.

Participants from across Harvard University included staff from Harvard University Archives, Preservation Services, Harvard University Information Technology Services, Countway Library, and Loeb Design Library, as well as members of the Harvard Library senior management team.

Article written by Harvard Library Communications.
Article published on March 16, 2016.



O Email! My Email! Our Fearful Trip is Just Beginning: Further Collaborations with Archiving Email

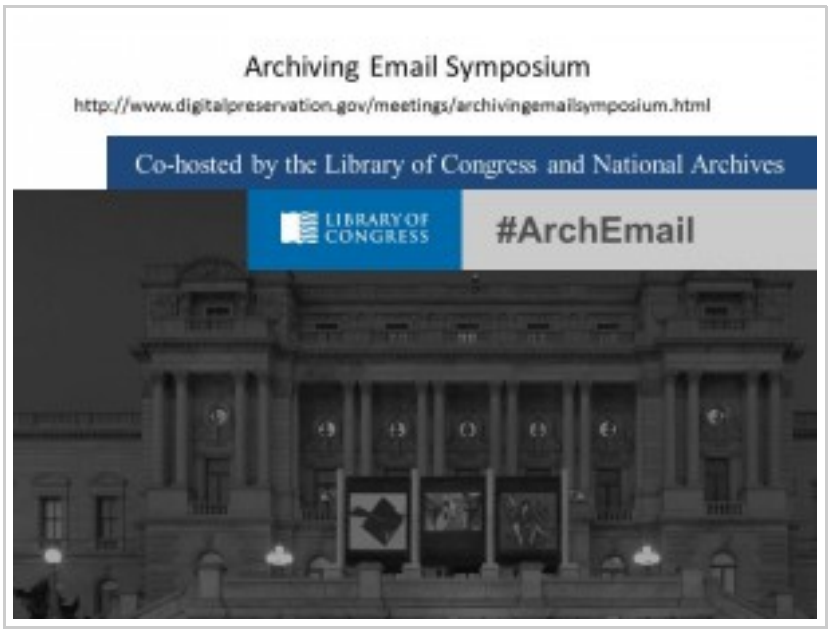
May 10, 2016 by Kate Murray

Apologies to Walt Whitman for co-opting the first line of his famous poem O Captain! My Captain! but solutions for archiving email are not yet *anchor'd safe and sound*. Thanks to the collaborative and cooperative community working in this space, however, we're making headway on the journey.

Email Archiving Stewardship Tools Workshop final panel. Franziska Frey, Christopher Prom, Glynn Edwards, Riccardo Ferrante, and Wendy Gogel. Photo courtesy of Kari Smith.

Email archiving as a distinct research area has been around a while but the discipline is still very much emergent. Stanford University Library, for example, has been working on acquiring and processing email from collections since 2010. ePADD's Glynn Edwards can trace her initial conversation on developing email archiving software with Smithsonian Institution Archives' Ricc Ferrante at the 2012 Society of American Archivists conference in San Diego and she agrees it is very gratifying to see the growth of support and interest, especially over the past year.

The Archiving Email Symposium (videos of the presentations are now available), hosted by the Library of Congress and the National Archives in June 2015, was one of the inspirations for the Email Archiving Stewardship Tools (Harvard EAST) workshop at Harvard Library on March 2-3, 2016. In addition to Harvard and the Library of Congress, participants for the workshop included the Smithsonian Institution Archives, Stanford University Libraries' ePADD project, MIT Institute Archives and Special Collections, University of Illinois Urbana-Champaign, Artefactual Systems and BitCurator Consortium.



The high-level goals of the two-day workshop, organized by Harvard's Wendy Gogel and Grainne Reilly, were community building, updating each other on current work, identifying and prioritizing gap areas and exposing the HL community to email-archiving efforts in the field at large. Just bringing the group together ticked off the first goal so we started the day with a mark in the win column.

Glynn Edwards summed up the mood in the room this way: "It was exciting to be part of the working group at Harvard sharing information about our various tools, processes, and needs and to begin conceptualizing a path of data and metadata through different tools contingent on their workflows. There was a lot of energy in the room and a willingness to work together to find ways to re-purpose metadata between tools and collaborate on building shared lexicons to assist with processing and discovery."

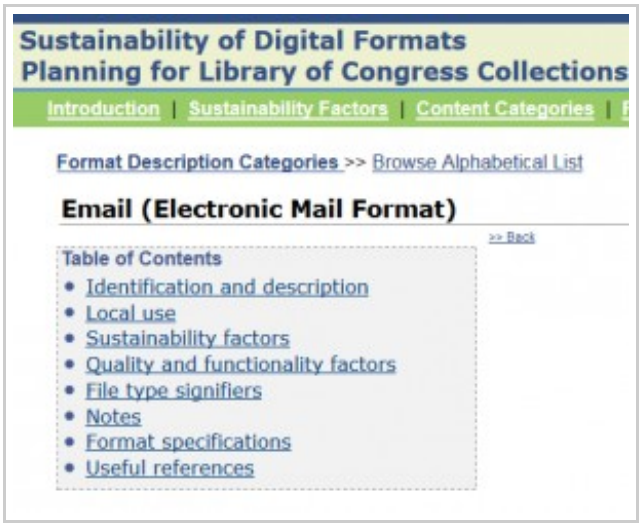
Edwards also found inspiration in Prom's statement that "email is one of the richest, one of the most revealing, if not **the** most revealing, of sources currently being generated." She goes on to say that "while correspondence has always been an important format in archival collections; email is often more – more immediate, more complex, more exposing. This is highlighted again on an almost weekly basis in breaking news – as the Governor's emails regarding Flint Michigan water crisis were released or emails and documents referred to as the Panama Papers were leaked."

Harvard's Widener Library. Photograph courtesy of Kate Murray

My personal interest is in the digital formats used for email messages and other personal information manager or PIM formats including calendaring, text and instant messages. As Prom indicated in the DPC Technology Watch Report Preserving Email (PDF), there's a convergence in the email archiving community around the MBOX family and EML as de facto preservation formats for email messages primarily because of two related factors: transparency and integration with toolsets.

The Library of Congress's Sustainability of Digital Formats website defines transparency, one of seven sustainability factors, as "the degree to which the digital representation is open to direct analysis with basic tools, including human readability using a text-only editor."

Native or normalized MBOX and EML files also can be used as access copies because they can be imported into a variety of email clients. It's no surprise then that these two plain text and very transparent formats, MBOX and EML, are integrated into popular email archiving tools and most modern email clients can import and export one or both of the formats. The Smithsonian Institution Archives' CERP toolset ingests MBOX-formatted messages before converting to XML, as will the still-in-development DArcMail (Digital Archive Mail System). The ePADD project developed at Stanford University Libraries also requires MBOX for ingest. Harvard University Libraries' Electronic Archiving System (EAS) ingests EML-formatted messages.



EML format description from LC's Sustainability of Digital Formats website

Harvard EAST workshop participants discussed some of the issues with these formats, including the lack of format validation tools and the challenges of working with formats, like MBOX, without documented standards.

Reflecting again on Whitman's poem, email archiving is still a work in progress and our voyage of discovery is nowhere near *closed and done*. However, projects like the Harvard EAST workshop move us all further along.

The MBOX format family from the Sustainability of Digital Formats website

Posted in: Digital Content, Partners and Collaboration, Tools and Infrastructure

Add a Comment »

Add a Comment

This blog is governed by the general rules of respectful civil discourse. You are fully responsible for everything that you post. The content of all comments is released into the public domain unless clearly stated otherwise. The Library of Congress does not control the content posted. Nevertheless, the Library of Congress may monitor any user-generated content as it chooses and reserves the right to remove content for any reason whatever, without consent. Gratuitous links to sites are viewed as spam and may result in removed comments. We further reserve the right, in our sole discretion, to remove a user's privilege to post content on the Library site. Read our Comment and Posting Policy.

Required fields are indicated with an * asterisk.

*** Name (no commercial URLs)**

*** Email (will not be published)**

*** Comment**

HARVARD
LIBRARY